COEFFICIENT OF INTRINSIC DEPENDENCE: A NEW MEASURE OF ASSOCIATION

A Dissertation

by

LI-YU DAISY LIU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2005

Major Subject: Statistics

COEFFICIENT OF INTRINSIC DEPENDENCE: A NEW MEASURE OF ASSOCIATION

A Dissertation

by

LI-YU DAISY LIU

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Tailen Hsing (Chair of Committee)	
Naisyin Wang (Member)	Edward R. Dougherty (Member)
Ruzong Fan (Member)	Michael T. Longnecker (Interim Head of Department)

May 2005

Major Subject: Statistics

ABSTRACT

Coefficient of Intrinsic Dependence:
A New Measure of Association. (May 2005)
Li-yu Daisy Liu, B.S., National Taiwan University,
M.S., National Taiwan University
Chair of Advisory Committee: Dr. Tailen Hsing

To detect dependence among variables is an essential task in many scientific investigations. In this study we propose a new measure of association, the coefficient of intrinsic dependence (CID), which takes value in [0,1] and faithfully reflects the full range of dependence for two random variables. The CID is free of distributional and functional assumptions. It can be easily implemented and extended to multivariate situations.

Traditionally, the correlation coefficient is the preferred measure of association. However, it's effectiveness is considerably compromised when the random variables are not normally distributed. Besides, the interpretation of the correlation coefficient is difficult when the data are categorical. By contrast, the CID is free of these problems. In our simulation studies, we find that the ability of the CID in differentiating different levels of dependence remains robust across different data types (categorical or continuous) and model features (linear or curvilinear). Also, the CID is particularly effective when the dependence is strong, making it a powerful tool for variable selection.

As an illustration, the CID is applied to variable selection in two aspects: classification

and prediction. The analysis of actual data from a study of breast cancer gene expression is included. For the classification problem, we identify a pair of genes that best classify a patient's prognosis signature, and for the prediction problem, we identify a pair of genes that best relates to the expression of a specific gene.

To my parents

ACKNOWLEDGEMENTS

I wish to express the deepest appreciation to my advisor, Dr. Tailen Hsing. Dr. Hsing has trained me as an independent and professional researcher. Through my years of Ph.D. study, he has been greatly supportive with his trust, patience and encouragement.

I also wish to thank Drs. Edward R. Dougherty and Naisyin Wang. They constantly inspire me by providing the insights into the ways of approaching different research fields.

Drs. Marcel Brun and Ulisses Braga-Neto kindly provided me computational resources and guidance about analysis of realistic microarray data during their stay in Genomic Signal Processing Laboratory at Texas A&M University. My special thanks are given to Dr. Marcel Brun for assisting me with producing some graphical results in this dissertation.

I am particularly grateful to my parents and my family for their love and comfort. My boyfriend, Jen-Hung, has consoled me spiritually and walked me through obstacles in the past year. Without their support, I could not have come this far.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

To detect and properly measure dependence among variables is an essential task in scientific investigations. Some famous instances in earlier ages include researchers in biology and anthropology describing the relationship between two or more characters of the same individual (Galton, 1889), economists constructing business barometers by inspecting the correlation between wholesale prices and a series of statistics indicating business conditions (Persons, 1916), and medical researchers examining the effectiveness of vaccination through observing the development of scars (Macdonell, 1902).

With science pregressing at the current pace, measuring dependence among variables in experiements is increasingly more important. The study of genetic regulation theories of protein synthesis, for example, typically contains a substantial element that addresses the magnitude of dependence of different genes and proteins. This is especially true after the rapid development of modern high throughput technologies (e.g. microarrays) for gene expression studies (Ermolaeva et al., 1998; Zhang, 1999). In the analysis of microarray expression data, an appropriate association index will be useful for understanding coordinated expression patterns across arrays (e.g. Eisen et al., 1998; Getz et al., 2000; Bergmann et al., 2004), relating disease phenotypes to gene expression patterns (e.g. Golub et al., 1999; Alon et al., 1999; van't Veer et al., 2002; Adryan and Schuh, 2004; Antonov et al., 2004), classifying genes according to

---

The format and style follow that of *Biometrics*.

their functional roles (e.g. Brown et al., 2000; Adryan and Schuh, 2004; Szabo et al., 2004), and sometimes simply reducing the dimensionality of objects under study (e.g. Carreira-Perpiñán, 1997; Broët et al., 2004; Chen et al., 2004).

Though it is now clear that the need to have suitable dependence measures is crucial in practical problems, the development of such measures did not receive sufficient attention until the late 19th century (Mari and Kotz, 2001) — In 1888, Sir Francis Galton conceptually defined the correlation coefficient by introducing a two-dimensional diagram plotting the sizes of daughter peas against the sizes of mother peas; the well-known mathematical framework for the correlation coefficient was derived by Karl Pearson a few years later (Stigler, 1986; Rodgers and Nicewander, 1988). It was found soon that the application of Galton-Pearson's correlation coefficient applies to a pair of continuous-valued variables, and the format of the data may not meet this requirement. This led to the unfolding of the methods for nominal and ordinal scales (Wherry, 1984). Spearman's ρ (1904) was developed to deals with ranks; Richardson and Stalnaker's point-biserial correlation coefficient (1933) permitted finding the relationship between a dichotomy and interval scores; Pearson proposed the coefficient of contingency which was derived from his goodness-of-fit Chi-squared test (Goodman and Kruskal, 1979). More discussion and references are included in Chapter III in this dissertation.

While other measures of association dealt with the difficulties caused by the data character, the (Galton-Pearson's) correlation coefficient still faces other limitation. First, the correlation coefficient is most effective under normality while real data often violates normality (Kat, 2003; Micceri, 1989; Neuhäuser and Lam, 2004; Rasmussen, 1986; Stigler, 1973; Thomas et al., 2001; Wilcox, 1990). To define correlation in terms of covariance can somewhat relieve this problem but the normal assumption is still essential to test hypotheses and to construct confidence intervals (Brutlag,

1998). Secondly, the correlation coefficient is really designed to describe the linear relationship (Chen and Popovich, 2002; Draper and Smith, 1998; Granger et al., 2004). Should a nonlinear pattern occur, one can either transform the data or employ a more sophisticated approach tailored for the problem, such as polynomial regression, although justifying such actions is usually problematic. (Greenland, 1996).

Last but not least, the correlation coefficient cannot properly explain the cause-effect relationship. In the computation of correlation coefficient, the activity of neither variables should be viewed as the result of the alternation of the other. One typical reason for which the correlation coefficient cannot infer causal effects is the possibility of influence from a third variable. Nevertheless, the confusion about correlation and causality still can be frequently observed in the literature (Chen and Popovich, 2002). The correlation coefficient's inability to interpret causal effects does not spare the thoughts for wanting to understand causal effects. In fact, the contrary is true. Pedhazur and Schmelkin (1991) argue that a causal framework is indispensable in research and in practice when we attempt to explain phenomena. James et al. (1982) also suggest that an understanding of causation is helpful for us to be able to make inference.

With the above concerns in mind we ask: What constitutes some of the basic criteria of a measure of dependence which makes good statistics sense and is flexible and powerful enough for analyzing a wide variety of data? We answer with the following:

(C1) The measure is model-free in the sense that no distributional or functional assumptions are placed on the variables; it is also invariant under monotone transformations of the marginals. This allows us to estimate the measure from data without having to verify model assumptions and/or make transformations.

This consideration is obviously important for data which the distributional properties are not well understood.

(C2) The measure conventionally takes values between 0 and +1 inclusive. It is +1 in the case of "complete association", and is zero in the case of independence.

(C3) The measure can fully differentiate different levels of dependence. For instance, the measure of dependence of a response variable on a predictor variable should become stronger if additional information is included in the prediction, or if the model is such that the response variable is functionally or stochastically more dependent on the predictor variable.

(C4) The measure takes causality into consideration and is not necessary to be symmetric. In other words, the dependence of X on Y may be different from the dependence of Y on X.

(C5) The measure can be easily and efficiently estimated from the data while measurement errors occur. It is also applicable to both continuous and categorical distributions and can be easily extended to multivariate distributions.

Note that our basic criteria do not require a dependence measure to reveal the nature or direction of dependence which we believe should be pursued separately.

The main goal of this dissertation is to introduce a new dependence measure that satisfies most if not all of the above criteria, and to illustrate how it can be used to solve a number of statistical problems. We call our new measure of dependence *the coefficient of intrinsic dependence*, or CID. The main motivating idea is that Y is strongly/weakly dependent on X if and only if the conditional distribution of Y given X is significantly/mildly different from the marginal distribution of Y. We measure

the difference by the normalized integrated squared distance so that the full range of dependence can be adequately reflected as numbers between 0 and 1.

The definition of CID will be given in Chapter II, and will be immediately followed by its estimation from the sample. The hypothesis test whether the target and predictor variable are independent usually is the major concern of many studies. Therefore, such tests based on CID will be constructed in the end of Chapter II. The comparison of CID with traditional correlation indices and other measures of dependence will be included in Chapter III. Four designs will be adopted in the experiments in order to investigate the performance of CID and other measures of dependence under linear and nonlinear circumstances. In Chapter IV, we will also the CID to do variable selection and illustrate how certain classification and prediction problems can be handled in that context. The analysis of a data set taken from a study of breast cancer (van de Vijver et al., 2002) will be demonstrated as well.

CHAPTER II

A NEW MEASURE OF ASSOCIATION

2.1 Introduction

The development of measure of associations can be initiated by looking up the definition of independence among variables. Two variables are called independence if the changes in the value of one have no effect on the value of the other. Casella and Berger (1990) provides more explicit definition for two one-dimensional random variables:

> Let (X, Y) be a bivariate random vector with joint pdf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called independent random variables if, for every $x \in \Re$ and $y \in \Re$,
>
> $$f(x, y) = f_X(x) f_Y(y).$$
>
> If X and Y are independent, the conditional pdf of Y given $X = x$ is
>
> $$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y),$$
>
> regardless of the value of x. It is equivalent to the saying if X and Y are independent then
>
> $$F(y|x) = F_Y(y), \tag{2.1}$$
>
> where $F(y|x)$ and $F_Y(y)$ are the conditional and marginal cdfs of Y, respectively.

The above statements can be immediately extended to multivariate cases. In contrast, two random variables are associated if they are not independent. Sometimes the

variable X and Y are referred as "predictor" and "target", respectively, in the sense of regression-like expression. The Equation (2.1) suggests that the distinction between marginal cdf of Y and the conditional cdf of Y given X might resolve the measure of associations. This idea was barely explored even though a similar concept was adopted to develop methods of analysis of variance in factorial experiments (Cheng and Jones, 2004).

Generally speaking, the comparison between marginal and conditional distributions is under the scope of goodness-of-fit (GOF) type measurement. The GOF statistics are famous for testing the hypothesis whether n observations have been drawn from a certain population. The Cramér-von Mises statistic, W^2, is among the most popular GOF statistics:

$$W^2 = n \int_{-\infty}^{\infty} \{G_n(x) - G(x)\}^2 dG,$$

where $G(\cdot)$ is a known distribution function from which one believes the sample is obtained and $G_n(\cdot)$ is the empirical distribution function. Observe that W^2 measures the squared discrepancy between $G(\cdot)$ and $G_n(\cdot)$. Some GOF statistics extend W^2 by adding a weight function in front of the squared portion. For instance, the Anderon-Darling statistic has the weight function $\{G(x)[1-G(x)]\}^{-1}$. For detailed information about GOF statistics, we refer to D'Agostino and Stephens (1986).

2.2 Coefficient of Intrinsic Dependence

Having the intention of comparing marginal and conditional cdfs as described in Equation (2.1), we begin with imitating the component in W^2:

$$\int_{-\infty}^{\infty} \{F(y|x) - F_Y(y)\}^2 dF_Y(y). \tag{2.2}$$

Out of consideration of different values of x, we take expectation over the random variable X and revise Equation (2.2):

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathrm{E}\{F(y|x) - F_Y(y)\}^2 dF_Y(y) \\
&= \int_{-\infty}^{\infty} \mathrm{E}\{P(Y \le y|X) - P(Y \le y)\}^2 dF_Y(y) \\
&= \int_{-\infty}^{\infty} \mathrm{Var}\{P(Y \le y|X)\} dF_Y(y) \\
&= \int_{-\infty}^{\infty} \mathrm{Var}\{\mathrm{E}(I(Y \le y)|X)\} dF_Y(y) \qquad (2.3)
\end{aligned}$$

where $I(A)$ is an indicator function, which is 1 if A is true and 0 otherwise. Yet the value manipulated by (2.3) can be any real number between 0 and infinity, it is more straightforward to have a measure of association taking a value between 0 and 1 for the convenience of illustration. We achieve the attempt by including a denominator in Equation (2.3). According to variance decomposition,

$$\mathrm{Var}\{I(Y \le y)\} = \mathrm{Var}\{\mathrm{E}[I(Y \le y)|x]\} + \mathrm{E}\{\mathrm{Var}[I(Y \le y)|x]\}$$

Hence, a new measure of association — coefficient of intrinsic dependence (or CID) — is proposed to be

$$\mathrm{CID}(Y|X) = \frac{\int_{-\infty}^{\infty} \mathrm{Var}\{\mathrm{E}[I(Y \le u)|x]\} dF_Y(u)}{\int_{-\infty}^{\infty} \mathrm{Var}\{I(Y \le v)\} dF_Y(v)}.$$

If Y is a continuous variable, the formula of CID can be alternatively written as

$$\mathrm{CID}(Y|X) = \frac{\int_0^1 \mathrm{Var}\left[\mathrm{E}[I(F_Y(Y) \le z)|x]\right] dz}{\int_0^1 \mathrm{Var}[I(F_Y(Y) \le w)] dw}.$$

We hereby provide detailed discussion about CID for continuous and discrete targets. Each case are demonstrated in two examples while their derivations are shown in Appendix.

2.2.1 CID for Continuous Targets

If the target variable, Y, is a continuous random variable then

$$\mathrm{E}[I(F_Y(Y) \leq u)] = P(F_Y(Y) \leq u) = u,$$

and

$$\mathrm{Var}[I(F_Y(Y) \leq u)] = P(F_Y(Y) \leq u) = u(1-u).$$

Therefore, the denominator of CID is

$$\int_0^1 u(1-u)du = 1/6.$$

With similar argument, the numerator of CID can be shortened as

$$\int_0^1 E[P^2(Y \leq F^{-1}(u)|X)]du - \frac{1}{3}.$$

So that, for a continuous Y,

$$\mathrm{CID}(Y|X) = 6\int_0^1 E[P^2(Y \leq F^{-1}(u)|X)]du - 2.$$

Example 2.1. Suppose X and Y are from a bivariate normal distribution with correlation ρ. It has been proven by Hsing et al. (2005) that

$$\mathrm{CID}(Y|X) = 6\sum_{k=1}^{\infty} \frac{\rho^{2k}}{k!} \int_{-\infty}^{\infty} \left(\phi^{(k-1)}(u)\right)^2 \phi(u)du,$$

where ϕ is the standard normal pdf. Note that $\mathrm{CID}(Y|X)$ is a strictly increasing function of $|\rho|$. □

Example 2.2. Let X and Y be taken from a exponential-gamma conjugate family:

$$X \sim gamma(\alpha, \beta); \quad Y|X \sim exp(x).$$

Then

$$\mathrm{CID}(Y|X) = 6\int_0^1 \left[2(1-u)^{-1/\alpha} - 1\right]^{-\alpha} du - 2,$$

which is fully determined by the shape parameter α. Figure 1, for instance, displays the motion of standard gamma pdf's (i.e. $\beta = 1$) with different values of α. The pdf of standard gamma is flatter as α gets larger, which results in the deviation of a conditional distribution from the marginal distribution of Y. □

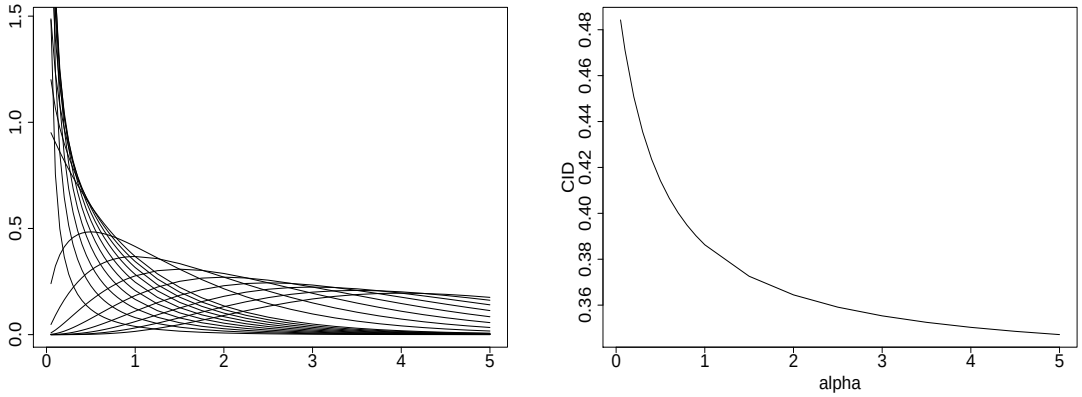


Figure 1: The plot on the left displays standard gamma pdf's. The pdf's, from left to right, have a fixed scale parameter $\beta = 1$ and the shape parameters α from 0.05 to 1 with 0.05 increment and from 1.25 to 5 with 0.25 increment. When α is less than or equal to 1, the pdf is concave; when α is greater than 1, the pdf is convex. The right-hand plot indicates the value of CID by a given value of α.

2.2.2 CID for Discrete Targets

A discrete target, Y, can be one of nominal, ordinal, or discrete interval variables. Without loss of generality, we assume Y takes values from 1 to m, where m is a positive integer or infinity. We firstly introduce the notations:

$$\begin{aligned} p_i &= P(Y = i); \\ p_i^+ &= \sum_{k=1}^{i} P(Y = k); \\ q_i^+(x) &= \sum_{k=1}^{i} P(Y = k|x). \end{aligned}$$

The denominator of CID is derived as

$$\sum_{i=1}^{m} p_i p_i^+ (1 - p_i^+).$$

The numerator appears to be

$$\sum_{i=1}^{m} p_i p_i^+ (1 - p_i^+) - \sum_{j=1}^{m} p_j \mathrm{E}[q_j^+(X)(1 - q_j^+(X))].$$

In summary, when Y is discrete,

$$\mathrm{CID}(Y|X) = 1 - \frac{\sum_{j=1}^{m} p_j \mathrm{E}[q_j^+(X)(1 - q_j^+(X))]}{\sum_{i=1}^{m} p_i p_i^+ (1 - p_i^+)}.$$

Example 2.3. If Y is binary then CID has a form pretty much alike with so-called coefficient of determination in regression analysis:

$$\mathrm{CID}(Y|X) = \frac{\mathrm{Var}[\mathrm{E}(Y|X)]}{\mathrm{Var}(Y)}.$$

In particular, when X is also binary,

$$\mathrm{CID}(Y|X) = r^2,$$

where r is the correlation coefficient. Another special case assumes that

$$X \sim beta(a, b) \text{ and } Y|X \sim Bernoulli(x).$$

Then

$$\mathrm{CID}(Y|X) = \frac{1}{a + b + 1}.$$

However, there is no such explicit formula for the cases of non-binary Y. □

Example 2.4. Consider the data represented in Table 1. We have $\mathrm{CID}(Y|X) = 0.019$ and $\mathrm{CID}(X|Y) = 0.053$. This example has shown the asymmetric property of CID.□

Table 1: One example of two-way contingency tables.

		X		
Prob.		1	2	Total
	1	0.09	0.09	0.18
Y	2	0.25	0.07	0.32
	3	0.38	0.12	0.50
Total		0.72	0.28	1.00

2.3 Properties

Several properties of CID are illustrated in this section.

1. CID requires minimal distributional assumptions. It is also invariant under variable transformations.

2. The causal relationship between variables is taken into account by asymmetric property of CID. That is, $\mathrm{CID}(Y|X)$ is not necessary to be the same as $\mathrm{CID}(X|Y)$. Example 2.4 in Section 2.2.2 had made this point.

3. CID always has a value between 0 and 1. If two random variables X and Y are independent to each other, then $\mathrm{CID}(Y|X) = \mathrm{CID}(X|Y) = 0$. In the other hand, $\mathrm{CID}(Y|X) = \mathrm{CID}(X|Y) = 1$ when X and Y are fully dependent. A number of simulations in Chapter III demonstrate that $\mathrm{CID}(Y|X)$ indeed increases if the model of X, Y changes in such a way that X asserts a larger influence on Y functionally or stochastically.

4. CID is ready to be implemented in different occasions, such as numerical, categorical, or multivariate cases, by inserting appropriate distribution functions.

2.4 Estimation

To estimate CID$(Y|X_1,\ldots,X_k)$ from data, for simplicity of notation, assume that $k=1$ and we observe the data $Y_i, X_i, 1 \le i \le n$. If $k>1$ then replace the univariate cdf's by multivariate ones. Suppose that Y's take $m \le n$ distinct values, and $\hat{p}_j$ is the observed proportion of the jth value; $1 \le j \le m$. The denominator of CID is estimated by

$$\sum_{j=1}^{m} \hat{p}_j \hat{p}_j^+ (1-\hat{p}_j^+), \tag{2.4}$$

where

$$\hat{p}_j^+ = \sum_{k=1}^{j} \hat{p}_k.$$

In particular, if all values of Y's are distinct then (2.4) is simply

$$\frac{1}{6} - \frac{1}{6n^2} \approx \frac{1}{6} \text{ for large n.}$$

The numerator of CID is more complicated. In the cases of continuous predictors, the bining process is necessary to estimate the conditional distribution. Let $A_l, 1 \le l \le a$, be a partition of the real line and n_l be the number of $X_i \in A_l$. Firstly define

$$\hat{q}_{kl} = \frac{\sum_{i=1}^{n} I(Y_i = k, X_i \in A_l)}{\sum_{i=1}^{n} I(X_i \in A_l)}, \quad \text{and } \hat{q}_{jl}^+ = \sum_{k=1}^{j} \hat{q}_{kl}.$$

Observe that $\hat{q}_{kl}$'s represent the conditional proportions. Then the estimate of numerator of CID is

$$\sum_{j=1}^{m} \hat{p}_j \hat{p}_j^+ (1-\hat{p}_j^+) - \sum_{j=1}^{m} \hat{p}_j \sum_{l=1}^{a} \frac{n_l}{n} \hat{q}_{jl}^+ (1-\hat{q}_{jl}^+), \tag{2.5}$$

with no doubt that the first part is the estimate of the denominator of CID. The choice of A_l is clearly a delicate issue. Let us assume for convenience that

$$\sum_{i=1}^{n} I(X_i \in A_l) = \frac{n}{a},$$

where a is the number of partitions. As a rule of thumb, a can be determined in such way that each partition contains more than five observations. Example 2.5 demonstrates the effect of bin size on CID estimation for different sample size

Example 2.5 Consider the following model:

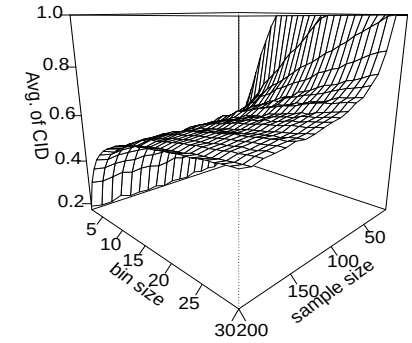
$$Y = X + X^2 + .5Z,$$

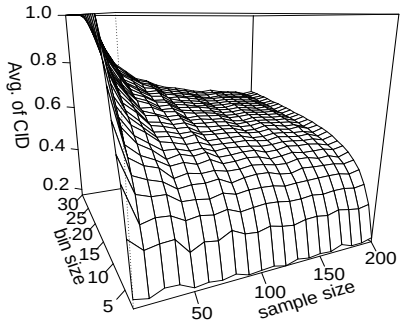
where X and Z are independent standard normals. Figure 2 summarizes the results of CID$(Y|X)$ estimation together with the MSE, variance and bias for different combinations of bin numbers and sample sizes based on 20 simulations. The set of plots are for sample sizes 10 – 200. One can see that CID estimator works very well for sample sizes as small as 50 and a wide range of bin numbers. Based on a variety of models, we are comfortable that these conclusions hold quite generally so long as both X and Y are one-dimensional. If X is multi-dimensional and the ith dimension uses a_i partitions, then X can be viewed as one-dimensional with bin size $\prod_i a_i$. It implies that the required sample size grows exponentially along with the increase of dimensionality and the estimation suffers so-called the "curse of dimensionality". □

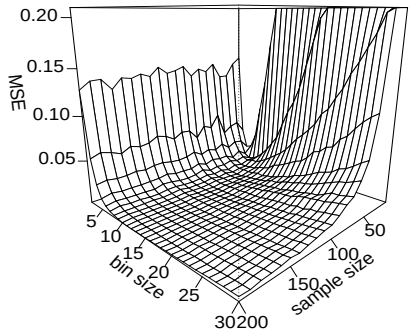
2.5 Hypothesis Tests of Independence

2.5.1 Introduction

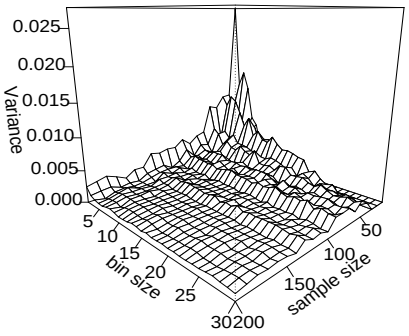
Mostly researchers are interested in examining if variables are independent to each other. For two categorical variables, Pearson's χ^2-statistic is capable of inspecting the absence of relationship. For two continuous variables, the hypothesis tests whether the population correlation coefficient, ρ, is different from 0 are usually enacted. However, the correlation coefficient demands assertion of strong assumptions. Otherwise, the inference may be misleading. In this chapter, we offer an alternative of hypothesis tests based on CID. A larger value of estimated CID indicates less evidence of existence



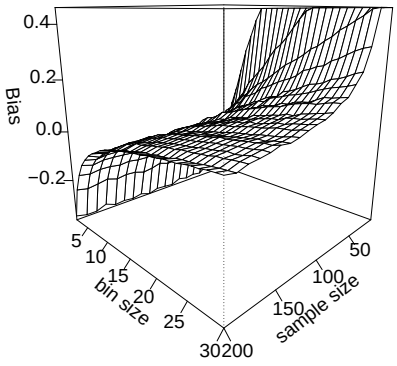




MSE



Variance



Bias

Figure 2: The two plots in the first row are CID values for different combination of bin and sample sizes from two perspectives. They are followed by the plots of MSE, variance, and bias.

of association. It is required to obtain the sampling distribution of estimates of CID under independence, which can be achieved by simulations. Furthermore, we compare the tests based on CID with those based on correlation.

2.5.2 Sampling Distributions

Let X and Y denote predictor and target variables, respectively. Suppose Y is continuous. If X and Y are independent, the conditional distribution of Y has to act like the marginal distribution of Y regardless of the given value of X. In the other words, the conditional distribution of Y would alter very little under the permutation of observations of X while the observations of Y remain in place. The sampling distribution of CID is manipulated based this idea. One firstly specify the number of observations, n, and the number of categories of X, a. It is straightforward to generate a sample Y of size n from a uniform distribution. A sample X can be generated in a way that there are approximately the same number of observations in each of a possible categories. One CID$(Y|X)$ estimate can be obtained from the generated paired observations. Repeat permuting the labels of X and compute the estimated CID$(Y|X)$ until N statistics are obtained. These N statistics contribute to the portrait of sampling distribution of CID. Figure 3 presents the histogram of one simulated sampling distribution with $n = 50$, $N = 1000$, and $a = 7$. The estimate of $100(1-\alpha)\%$ quantile is consequently available for a level α hypothesis test of independence.

The sampling distribution of the correlation coefficient has been well established. Suppose the paired sample is bivariate normal distributed,

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \sim \quad t_{n-2},$$

when $\rho = 0$, where r is the sample correlation coefficient and t_{df} denotes the Students' t distribution with df degrees of freedom. However, for the sake of comparison,

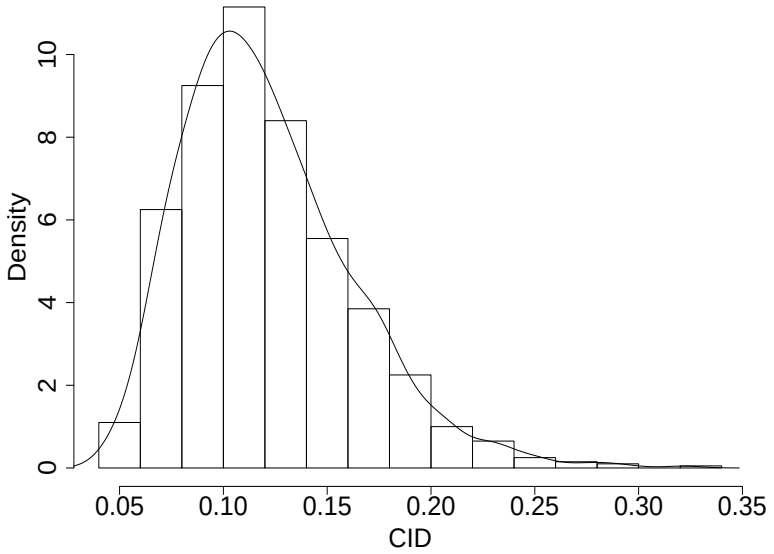


Figure 3: The plot displays the histogram and the estimated density function of sampling distribution of CID from one simulation when two variables are independent. There are 1000 CID estimates. Each estimate is computed from a paired sample of size 50. The bin size of predictor is set to be 7.

it is fairer to obtain the critical value for the hypothesis test of independence from simulations as well. There are N paired samples of size n are taken from a bivariate normal distribution with identity covariance matrix and the sample correlation coefficients are computed. Figure 4 displays one example of the quantile-quantile plot of theoretical t_{n-2} distribution and the N samples after proper transformation when $N = 1000$ and $n = 50$. A nice fitting to a 45 degree line suggests the suitability of the simulation.

2.5.3 Simulation Results of Independence Tests

Let's formally address the null and alternative hypotheses as bellow:

$$H_0 : Y \text{ does not depend on } X;$$

$$H_1 : Y \text{ depends on } X.$$

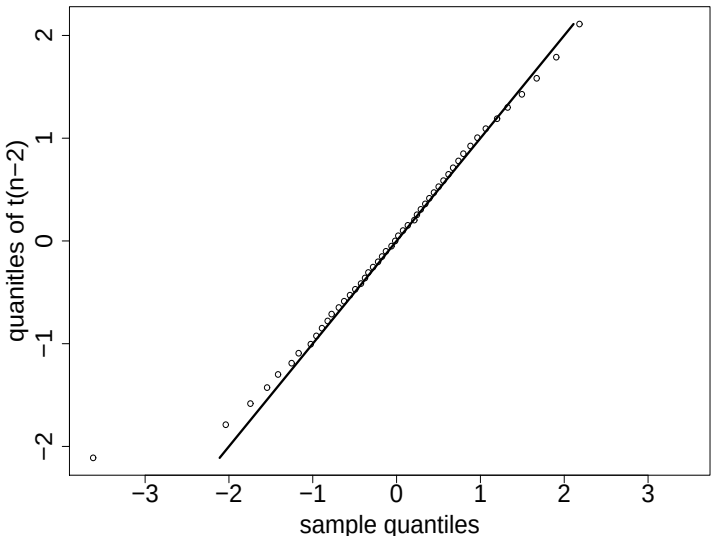


Figure 4: The quantiles of t distribution with degrees of freedom $n-2$ are plotted against the quantiles of $r_i\sqrt{n-2}/\sqrt{1-r_i^2}$, where r_i is the ith $(i = 1, \cdots, N)$ estimate of correlation coefficient from the sample of bivariate normal with $\rho = 0$. In this plot, $n = 50$ and $N = 1000$.

One tends to reject H_0 if the sample correlation or estimated CID is large. We compare the tests based on CID with those based on sample correlation while the cut-off points are determined by simulations. Three simulations are proceeded. In each simulation, we let $n = 50$, $N = 1000$, and $a = 7$.

In the first simulation we consider the model where (X, Y) are bivariate normal with correlation $\rho \in [0, 1]$. The power curves of level 0.05 CID-based and correlation-based tests are shown in Figure 5. Not surprisingly the correlation-based test is considerably more powerful. The correlation determines the distribution in the case of the normal and hence the correlation-based test is naturally optimal in this setting.

The other two simulations are designed to observe the power of tests based on either CID or correlation in more general class of models. Two models are applied in

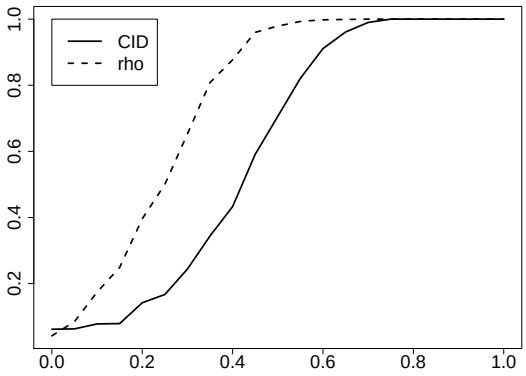


Figure 5: The samples are taken from bivariate normal with correlation $\rho \in [0, 1]$. Given ρ (x-axis), the plot shows the power of level 0.05 tests of independence based on CID (solid curve) and correlation (dashed curve).

the simulations:

$$Y = \rho X^2 + (1 - \rho)\epsilon \tag{2.6}$$

$$Y = \rho[\sin(2\pi(X - 1)/6) + \sin(2\pi(X - 1)/3)] + (1 - \rho)\epsilon \tag{2.7}$$

where X and ϵ are $N(0, 1)$. Due to nonlinear terms of X, the correlation does not adequately describe the dependence in the model. Figures 6 and 7 are the plots of the power curves for Model (2.6) and (2.7), respectively. The CID-based test appears to be more powerful than the correlation-based test in both cases.

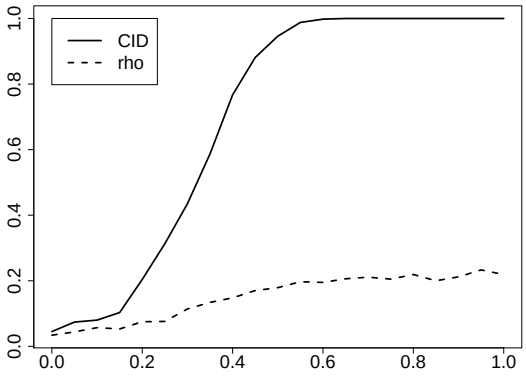


Figure 6: The samples are taken from Model (2.6) with $\rho \in [0, 1]$. Given ρ (x-axis), the plot shows the power of level 0.05 tests of independence based on CID (solid curve) and correlation (dashed curve).

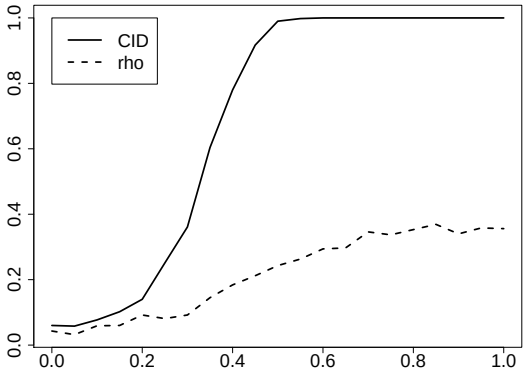


Figure 7: The samples are taken from Model (2.7) with $\rho \in [0, 1]$. Given ρ (x-axis), the plot shows the power of level 0.05 tests of independence based on CID (solid curve) and correlation (dashed curve).

CHAPTER III

COMPARISON OF MEASURES OF ASSOCIATION

3.1 Introduction

The strength of the associations is the principal concern in the study of relationship among variables, which are classified as continuous or discrete according to the possible values they theoretically possess. Variables that assume all possible values in certain interval are called continuous or interval variables. Discrete variables are those for which the set of all possible values is some discrete set of numbers; they are further classified as nominal or ordinal. Since discrete variables often identify categories into which population objects fall, they are also called categorical variables.

There are numerous measures currently available for different data classes. Due to the lack of such a statistic which can be universally used for different classes, the computation and interpretation of these traditional measures concerns the nature and the source of data. The most common way of quantifying the relationship between two continuous variables is the correlation coefficient under the assumption that two variables are bivariately normally distributed. Kendall's τ and Spearman's ρ statistics provide more options in general distributional settings. Correlation indices for two categorical variables or the mixture of two classes are also available (Chen and Popovich, 2002; Wherry, 1984). However, correlation indices aim to pick linear or monotone relationships. They may have problems to detect the existence of curvilinearity. Although performing transformations might lessen the impact, there is still the problem of choosing the appropriate transformation and the difficulty of further inferences (Draper and Smith, 1998).

Multinominal or hypogeometric settings for categorical variables allow more

flexibility of measuring associations. From three aspects statisticians usually evaluate the magnitude of dependence regardless of natural ordering of variables: the goodness of fit, the accuracy of prediction, and the reduction in variance. Additionally, if the categories for the variables remain natural order (e.g., high, median, or low education levels), there are methods that count the preciseness of predicting the ordering of categories. Some thorough descriptions and summaries for these statistics are available in Goodman and Kruskal (1979), Liebetrau (1983), and Agresti (1990).

In this chapter, we compare CID with traditional indices. Definitions of correlation indices and other measures of association are briefly described in Section 3.2. The simulation settings and results follow in Sections 3.3 and 3.4.

3.2 Traditional Measure of Association

3.2.1 Correlation Indices

The correlation is one of the most frequently used measure to describe the relationship between two variables. It is straightforward to manipulate and has a natural interpretation in bivariate normal distributions. Suppose X and Y are two real-valued random variables with finite variances. The correlation coefficient between X and Y is

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{V}(X)\mathrm{V}(Y)}},$$

where $\mathrm{Cov}(X,Y)$ is the covariance between X and Y defined by

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))] = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$$

and $\mathrm{V}(X)$, $\mathrm{V}(Y)$ denote the variances of X and Y. From the sample of n observations (X_i, Y_i), $1 \le i \le n$, ρ is estimated by

$$\hat{\rho}(X,Y) = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{(\sum X_i^2 - n\bar{X}^2)(\sum Y_i^2 - n\bar{Y}^2)}}, \tag{3.1}$$

where $\bar{X}$ and $\bar{Y}$ are sample means of X_i's and Y_i's. The correlation can be any real number between -1 and 1. If X and Y are independent then $\rho = 0$. In the cases of perfect linear dependence, we have $\rho = \pm 1$. Besides, the correlation is invariant under strictly increasing linear transformations but it is not invariant under nonlinear strictly increasing transformations. Another disadvantage of the correlation is that it is highly affected by extreme outliers.

Researchers usually turn to rank correlations to stretch the distributional assumption. There are two rank-based correlation indices, Spearman's ρ (denoted by ρ_s in this dissertation to be distinguished from the correlation coefficient) and Kendall's τ statistic. Let $\tilde{X}_i$ and $\tilde{Y}_i$ be the rank of X_i and Y_i. The ρ_s computes the correlation on the ranks of two random variables:

$$\hat{\rho}_s(X, Y) = \hat{\rho}(\tilde{X}, \tilde{Y}),$$

where $\hat{\rho}$ is the usual sample correlation. Kendall's τ accounts more generally the level of concordance between two random variables:

$$\hat{\tau}(X, Y) = \frac{1}{\binom{n}{2}} \sum_{i<j} sign[(X_i - X_j)(Y_i - Y_j)].$$

Both ρ_s and τ take real values between -1 and 1. They are distribution-free and robust against outliers.

The correlation coefficient, however, is originally developed for accessing the dependence between two continuous variables. To fulfill the inquiries of association measures for two discrete variables or the mixture of continuous and discrete variables, other correlation indices ought to be established. In the sense of generalizing the correlation, it is relatively easier to handle a dichotomous or a ordinal variable than to handle a multichotomous variable. For a dichotomous variable, one can manually code the data by either 0 or 1 and proceeds the computations of correlation. In fact,

the correlation remains the same value no matter what numbers are assigned to the categories of a dichotomous variable. From which two special cases of ρ are derived: the phi coefficient (ϕ) for two dichotomous variables and the point-biserial correlation (r_{pb}) for one dichotomous and one interval variable. If the variable has ordinal scales, data ranks can be utilized. It names two special cases of correlation, r_{RI} for ordinal versus interval variable and r_{DR} for ordinal versus dichotomous variable. With similar argument, Spearman's ρ_s can be adopted for the cases of two ordinal variables.

When encountering a multichotomous variable, there are no specific labels designated for the correlation. Suppose one variable is multichotomous and the other is interval, ordinal or dichotomous, Wherry (1984) suggested a dummy variable can be created for each class of the multichotomy by assigning the mean score of the members of that class on the other variable. These statistics are denoted by η, r_{MR}, or r_{MD}, depending on whether the other variable is continuous, ordinal, or dichotomous. However, the tactic of creating a pseudo variable breaks down while both variables are multichotomous. Some measures we bring forward in the next section are possible alternatives.

3.2.2 Other Measures of Association for Categorical Variables

A contingency table helps us understand the correspondence between two categorical variables. Suppose two categorical variables, X and Y, have A and B levels, respectively. Table 2 provides an example what a contingency table looks like. The cells of the table contain frequency counts of AB possible outcomes. We denote

$n_{ab} =$ the count of incidents that X takes the $a$the level and Y takes the bth level;

$$n_{a\cdot} = \sum_{b=1}^{B} n_{ab}; \quad n_{\cdot b} = \sum_{a=1}^{A} n_{ab}; \quad n = \sum_{a=1}^{A}\sum_{b=1}^{B} n_{ab}.$$

Table 2: The contingency table of two categorical variable X and Y. Variable X has A levels and variable Y has B levels.

	Y				
X	1	2	$\cdots$	B	Total
1	n_{11}	n_{12}	$\cdots$	n_{1B}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2B}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A	n_{A1}	n_{A2}	$\cdots$	n_{AB}	$n_{A\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot B}$	n

Additionally, let

$$n_{m\cdot} = \max_a n_{a\cdot}; \quad n_{\cdot m} = \max_b n_{\cdot b}; \quad n_{mb} = \max_a n_{ab}; \quad n_{am} = \max_b n_{ab}.$$

The chi-square statistic, χ^2, is proposed to examine the existence of association:

$$\chi^2 = \sum_{a=1}^{A}\sum_{b=1}^{B} \frac{n_{ij} - n_{a\cdot}n_{\cdot b}/n}{n_{a\cdot}n_{\cdot b}}. \tag{3.2}$$

It is straightforward that a association measure imitates the idea of the chi-squared statistic. Three statistics are derived in this fashion, including Pearson's C, Tschuprow's T, and Cramér's V:

$$\begin{aligned} C &= \sqrt{\frac{\chi^2}{n+\chi^2}}, \\ T &= \sqrt{\frac{\chi^2/n}{\sqrt{(A-1)(B-1)}}}, \\ V &= \frac{\chi^2/n}{\min(A-1, B-1)}, \end{aligned}$$

where χ^2 is defined in Equation (3.2). All three statistics are symmetric and have values between 0 and 1.

Another type of measures considers the optimal way of prediction on one variable with or without taking account of the knowledge of the other variable. It is the

proportion of deduction for estimation errors used to describe the level of dependence. Goodman-Kruskal's λ was firstly proposed based on this idea:

$$\lambda = \frac{\sum_a n_{am} + \sum_b n_{mb} - n_{\cdot m} - n_{m\cdot}}{2n - n_{\cdot m} - n_{m\cdot}}.$$

It is worthwhile to mention that the recently defined measure CoD (Dougherty et al., 2000) is motivated by an analogous idea using a variety of loss functions. The measures in this group have values between 0 and 1 but they are not necessarily symmetric.

The reduction in variance is also a potential indicator for association. Suppose that When two variables are more dependent, people can more precisely predict one variable by the other. So that little within-column or within-row variation can be expected while two variables are highly associated. To make this point, Goodman-Kruskal's concentration coefficient (τ^*) and Theil's uncertainty coefficient (U) each concern the proportional deduction for Gini concentration and entropy:

$$\begin{aligned} \tau^* &= \frac{n\sum_a\sum_b n_{ab}^2/n_{a\cdot} - \sum_b n_{\cdot b}^2}{n^2 - \sum_b n_{\cdot b}^2}; \\ U &= -\frac{\sum_a\sum_b n_{ab}\log(n\cdot n_{ab}/(n_{a\cdot}n_{\cdot b}))}{\sum_b n_{\cdot b}\log(n_{\cdot b}/n)}. \end{aligned}$$

Both measures take value between $[0, 1]$ and are asymmetric.

An ordinal variable is among one particular type of categorical variables. The natural ranking among categories remains directional even though the absolute distances are unknown. It is appropriate for such measures of association to take signed values indicating either agreement or disagreement between two variables. Besides, the permutation of rows and columns in contingency table would affect the results. Kendall's τ employs the same idea but it is impractical when there are ties. Several statistics follow its track: Kendall's τ_b, Goodman-Kruskal's γ, Stuart's τ_c, Somers's

d, and Wilson's e. Let Π_s, Π_d, and Π_t represent the probabilities of concordance, disconcordance, and ties, respectively. Then

$$\begin{aligned}
\tau_b &= \frac{n^2(\Pi_s - \Pi_d)}{\sqrt{(n^2 - \sum_a n_{a\cdot}^2)(n^2 - \sum_b n_{\cdot b}^2)}}; && (3.3)\\
\gamma &= \frac{\Pi_s - \Pi_d}{\Pi_s + \Pi_d}; && (3.4)\\
\tau_c &= \frac{\Pi_s - \Pi_d}{(m-1)/m}, \quad \text{where } m = \min(\alpha, \beta); && (3.5)\\
d &= \frac{n^2(\Pi_s - \Pi_d)}{n^2 - (\sum_a n_{a\cdot}^2 + \sum_b n_{\cdot b}^2)/2}; && (3.6)\\
e &= \frac{n^2(\Pi_s - \Pi_d)}{n^2 - \sum_a \sum_b n_{ab}^2}. && (3.7)
\end{aligned}$$

From the data Π_s, Π_d, and Π_t can be estimated by

$$\begin{aligned}
\hat{\Pi}_s &= 2\sum_a \sum_b \frac{n_{ab}}{n} \left(\sum_{a'>a} \sum_{b'>b} \frac{n_{a'b'}}{n} \right);\\
\hat{\Pi}_t &= \sum_a \left(\frac{n_{a\cdot}}{n}\right)^2 + \sum_b \left(\frac{n_{\cdot b}}{n}\right)^2 - \sum_a \sum_b \left(\frac{n_{ab}}{n}\right)^2;\\
\hat{\Pi}_d &= 1 - \hat{\Pi}_t - \hat{\Pi}_s.
\end{aligned}$$

The above estimates can be inserted into Equation (3.3) – (3.7) in order to obtain the sample statistics.

3.3 Simulation Setup

In Table 3 (a) we summarize all measures of associations introduced in Section 3.2. The appropriate usage of traditional statistics is subject to the nature of two variables while CID is globally applicable. Table 3 (b) shows that in our experiments we conveniently classify the indices in five groups.

- Group I are the correlation indices while both variables are categorical.
- Group II are the correlation indices for one discrete and one continuous variable.

- Group III are the correlation indices while both variables are continuous.
- Group IV are association measures for two categorical variables.
- Group V are association measures specifically for two ordinal variables.

Statistics in each group are performed on one acquainted linear model and one curvilinear model. In each simulation, the coefficient r ($0 \leq r \leq 1$) is used to control the strength of relationship; 0 for independence and 1 for full dependence.

In our experiments, the linear model performed on all of the statistics is simply the normal distribution. Given a level of dependence, r, a random sample is taken from the standard bivariate normal distribution which has r as its population correlation.

Design 1. A random sample of size 500 is taken from a standard bivariate normal distribution. One or both of the paired observations are categorized by theoretical quantiles of $N(0,1)$ for those indices not belonging to Group III.

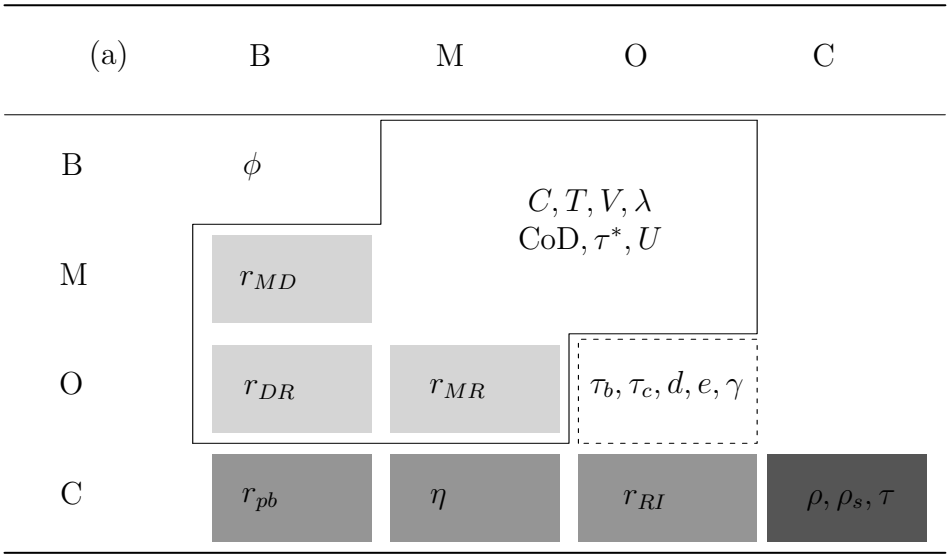
Samples from curvilinear models are developed case by case.

Design 2. This is for the cases of two continuous variables. A random sample, X, of size 500 is taken from $N(0,1)$ and let

$$Y_i = rX_i^2 + (1-r)\epsilon_i, \quad \epsilon_i\text{'s} \overset{iid}{\sim} N(0,1).$$

Design 3. One continuous variable (Y) and one discrete variable (X) are now of interest. This design is identical to the second one except X is classified into nc groups by theoretical quantiles of $N(0,1)$ and only the class labels (from 1 to nc) are recorded.

Table 3: Summary of association measures is provided. In (a) it displays the suitable measures(s) according to the types of two variables. (B = binary; M = multichotomous; O = ordinal; C = continuous.) It is inspired by Table 1.1 in Chen and Popovich (2002). The measures are divided into five groups as shown in (b).



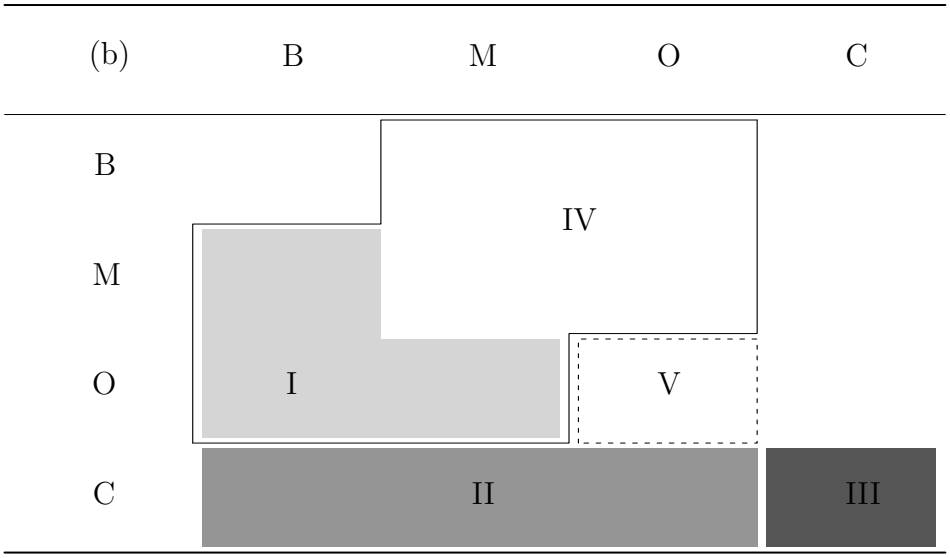


Table 4: The designs employed in each group of association measures are labelled by "⋆". Every group is evaluated in one linear and one curvilinear design.

	Index Group				
Design	I	II	III	IV	V
1	⋆	⋆	⋆	⋆	⋆
2			⋆		
3	⋆				
4	⋆			⋆	⋆

Design 4. Let both variables be categorical. We pre-determine the numbers of classes, nc and nr for two variables and one way of perfect prediction which is denoted by Λ. A simple random sample of size 500, X, is first drawn from 1 to nc. The corresponding value, Y_i, of X_i has chance r of taking the value determined by index Λ and $(1-r)$ of taking a simple random sample of size 1 from 1 to nr.

The groups of association methods matches the utilized simulation designs in Table 4. In each experiment, the sampling procedure iterates 50 times and the average estimates are documented.

3.4 Experimental Results

3.4.1 Comparison of CID with Correlation Indices

There are 50 samples of two continuous variables generated from Design 1 and 2. To estimate CID, we determine the number of partitions for the predictor to be 20. The results of comparison are shown in Figure 8. The linear model gradually promotes all four statistics from 0 to 1. Under the curvilinear model (Design 2), ρ, ρ_s and τ all yield a value approximate zero; only CID changes its value along with different levels of dependence. It is not surprising that correlation indices estimate the

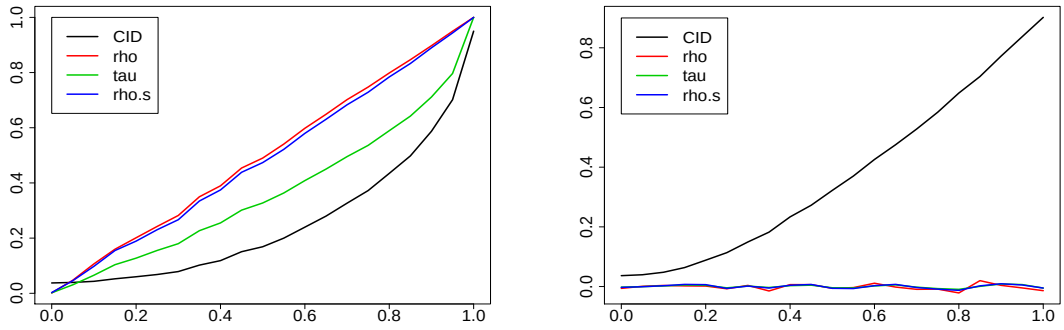


Figure 8: The correlation indices for two continuous variables, are compared with CID, including ρ (red), τ (green), and ρ_s (blue) statistics. The left plot is the experimental results when the linear model (Design 1) is considered. The plot on the right illustrates the results of curvilinear model (Design 2). The x-axes in two plots denote the degree of dependence, r.

Table 5: Two schemes of perfect prediction are adopted to generate the data of two discrete variables. The predictor and target variable are denoted by X and Y, respectively. Given the value of X, Λ_1 and Λ_2 explain the value of Y should be under absolute dependence.

Λ_1: $(nr, nc) = (2, 3)$

X	1	2	3
Y	1	2	1

Λ_2: $(nr, nc) = (3, 5)$

X	1	2	3	4	5
Y	1	2	3	2	1

correlation better than CID does if the sample is bivariately normally distributed. In terms of discrimination, however, more gradual increment of CID helps to distinguish association orders.

For two discrete variables, 50 samples are individually produced from Design 1 and 4. We consider two scenarios of (nr, nc) combinations, $(2, 3)$ and $(3, 5)$, due to the desire of different variable types (dichotomous or multichotomous) from different statistics. In Design 4, the corresponding perfect-prediction indices are Λ_1 and Λ_2 listed in Table 5. Figure 9 shows the simulation results. The comparison of CID with r_{DR} draws similar conclusion as that in the cases of two continuous variables. The other two correlation indices, r_{MR} and r_{MD}, seem to well estimate the level of

dependence in both linear and curvilinear designs. That is because they substitute X by the mean of Y given that specific value of X. It can be easily shown that r_{MD} and r_{MR} both take the value of $\sqrt{V(E(Y|X))/V(Y)}$ in different occasions. Particularly in Design 4,

$$r_{MD} = \sqrt{\frac{8r^2}{9-r^2}}, \quad r_{MR} = \sqrt{\frac{42r^2}{(5+r)(10-3r)}}, \quad \text{and} \quad r_{DR} = 0,$$

while

$$\text{CID}_{2\times 3} = \frac{8r^2}{9-r^2} \quad \text{and} \quad \text{CID}_{3\times 5} = \frac{54r^2(1+r)}{(5+r)^2(4-r)}.$$

Even though CID overall underestimate the true level of dependence, it once again has the advantage of differentiation due to perpendicular slope.

Finally we compare CID with three correlation indices for the mixture of interval and categorical variables. From Design 1 and 3, 50 samples are generated; $nc = 2$ and $nc = 5$ are under our study. The results are presented in Figure 10. Design 3 is senseless if $nc = 2$. However, CID shows stability while r_{pb} does not. If $nc = 5$, r_{RI} is instantly ruled out since it has constant zero all the time. As usual, CID is in favor to distinguish association levels greater than 0.7. The eta statistic might be used if differentiating lower level of dependence is of interest.

3.4.2 Comparison of CID with Other Measures of Association

To compare CID with other association measures illustrated in Section 3.2.2, the same trials described in Table 5 has been established. From each of Λ_1 and Λ_2, we simulate 50 samples and compute all estimates; Figures 11 and 12 are the results based on the two scenarios, respectively.

There are five measures have to assume ordinal variables, including τ_b, γ, τ_c, d, and e statistics. They are expected to act like regular correlation coefficient. When encountering linear model, these five statistics seem to be positively reciprocal to the

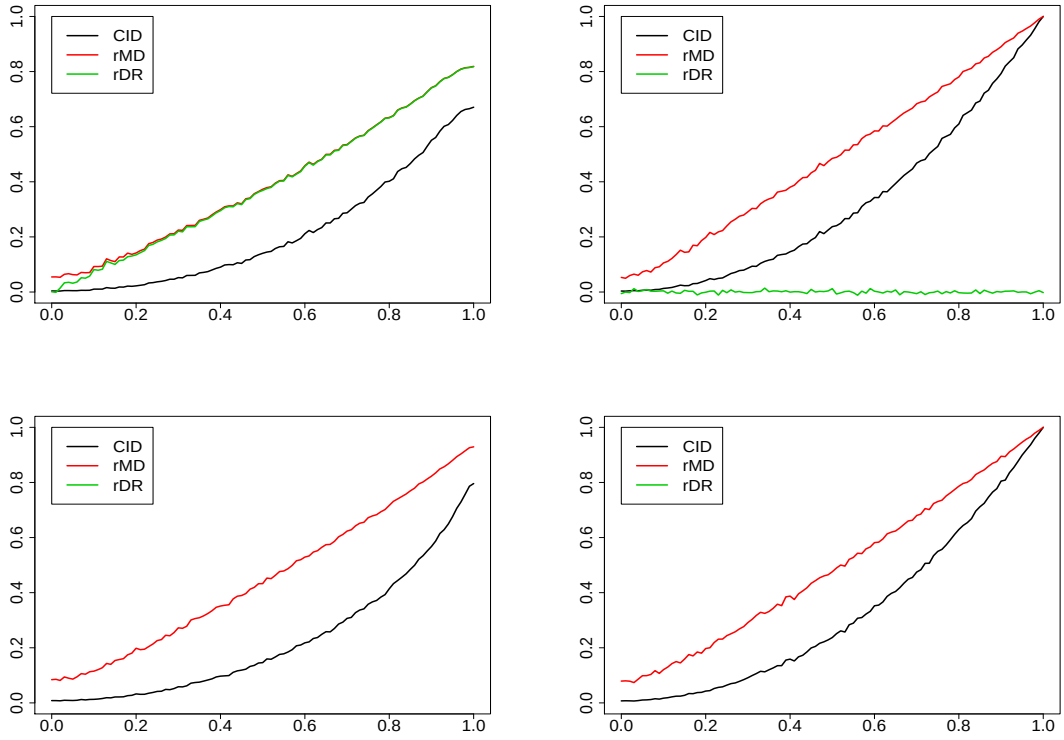


Figure 9: The correlation indices for two discrete variables are compared with CID, including r_{MD} for one dichotomous and one multichotomous variable, r_{DR} for one dichotomous and one ordinal variable, and r_{MR} for one multichotomous and one ordinal variable. r_{MD} and r_{DR} are tested on the data with 2-category target and 3-category predictor while r_{MR} are tested on the data with 3-category target and 5-category predictor. The left panel are the experimental results when the linear model (Design 1) is applied. The plots on the right panel illustrates the results of curvilinear model (Design 4). The x-axis in each plot denotes the specified degree of dependence, r.

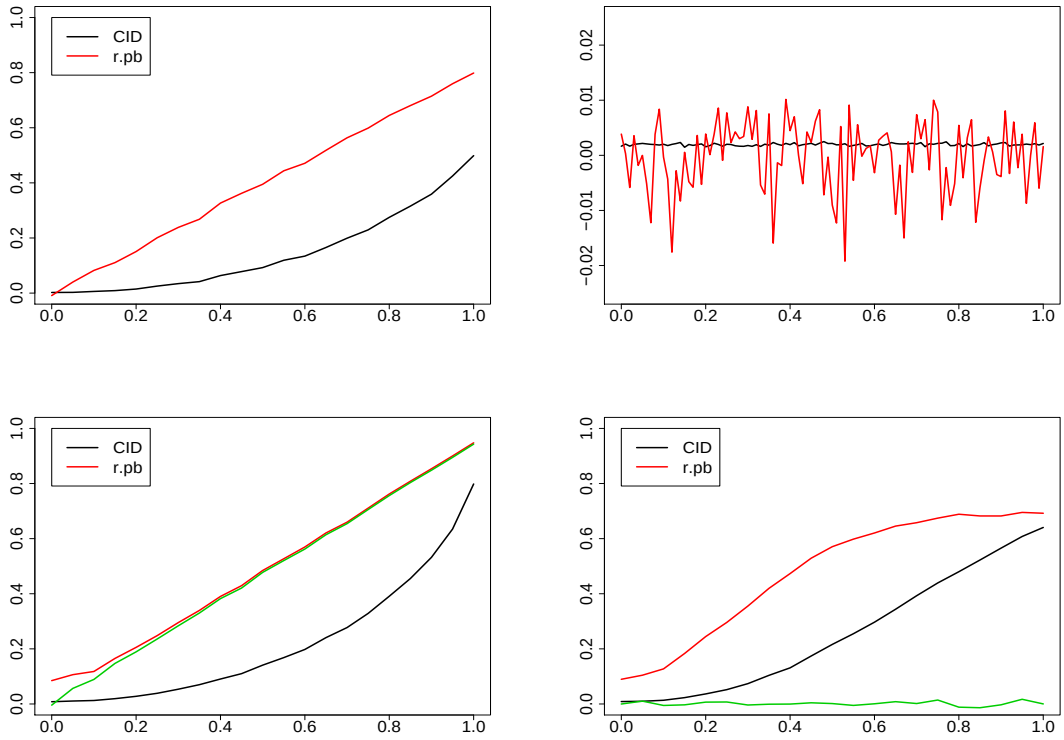


Figure 10: The correlation indices for one discrete variable and one interval variable are compared with CID, including r_{pb}, η and r_{RI} for the cases that the categorical variable is dichotomous, multichotomous, and ordinal, respectively. In our experiments, the predictor is set to be categorical to avoid possible effect from the choices of bin sizes. The data with two-category predictor is generated while comparing CID with r_{pb}. Otherwise, the predictor is generated to have five categories. The left panel are the experimental results when the linear model (Design 1) is applied. The plots on the right panel illustrates the results of curvilinear model (Design 3). The x-axis in each plot denotes the specified degree of dependence, r.

degree of association, r. However, they easily break down when linearity is no longer to be held.

Statistics among the other three categories remain some robustness in different model assumptions. Experimental results show that the estimates go up along with the increment of dependence level. One interesting fact is that CoD eventually achieves the value one but its path is relatively bumpy if the target variable has more than two possible levels. The results also reveal that CID is particularly analogous to Goodman-Kruskal's concentration coefficient (τ^*) and Theil's uncertainty coefficient (U). This is perhaps from different aspects they all concern the reduction in variance.

3.5 Discussion

We have reviewed various measures of association applicable to different types of data. Among all of these measures, the correlation coefficient is the most widely adopted statistic for continuous data. When the correlation coefficient is calculated to describe characteristics of a sample, it requires no distributional assumption. However, researchers are prevented from making definitive inferential conclusions (e.g., to conduct null hypothesis tests) without bivariate the normal assumption. Also the correlation coefficient is not invariant under monotone transformations, and this consideration gives rise to two alternative methods, Spearman's ρ (denoted by ρ_s in this dissertation) and Kendall's τ. They measure the relationship on a rank basis.

On the other hand, categorical variables are frequently present in the studies. It is sometimes the difficulty of defining metrics to the categories that obstructs the direct application of correlation coefficient. For a dichotomous variable, arbitrary numbers can be assigned for the two categories due to the invariant property of correlation coefficient under transformations for binary variables (ϕ, r_{DR}, and r_{pb}). For an ordinal

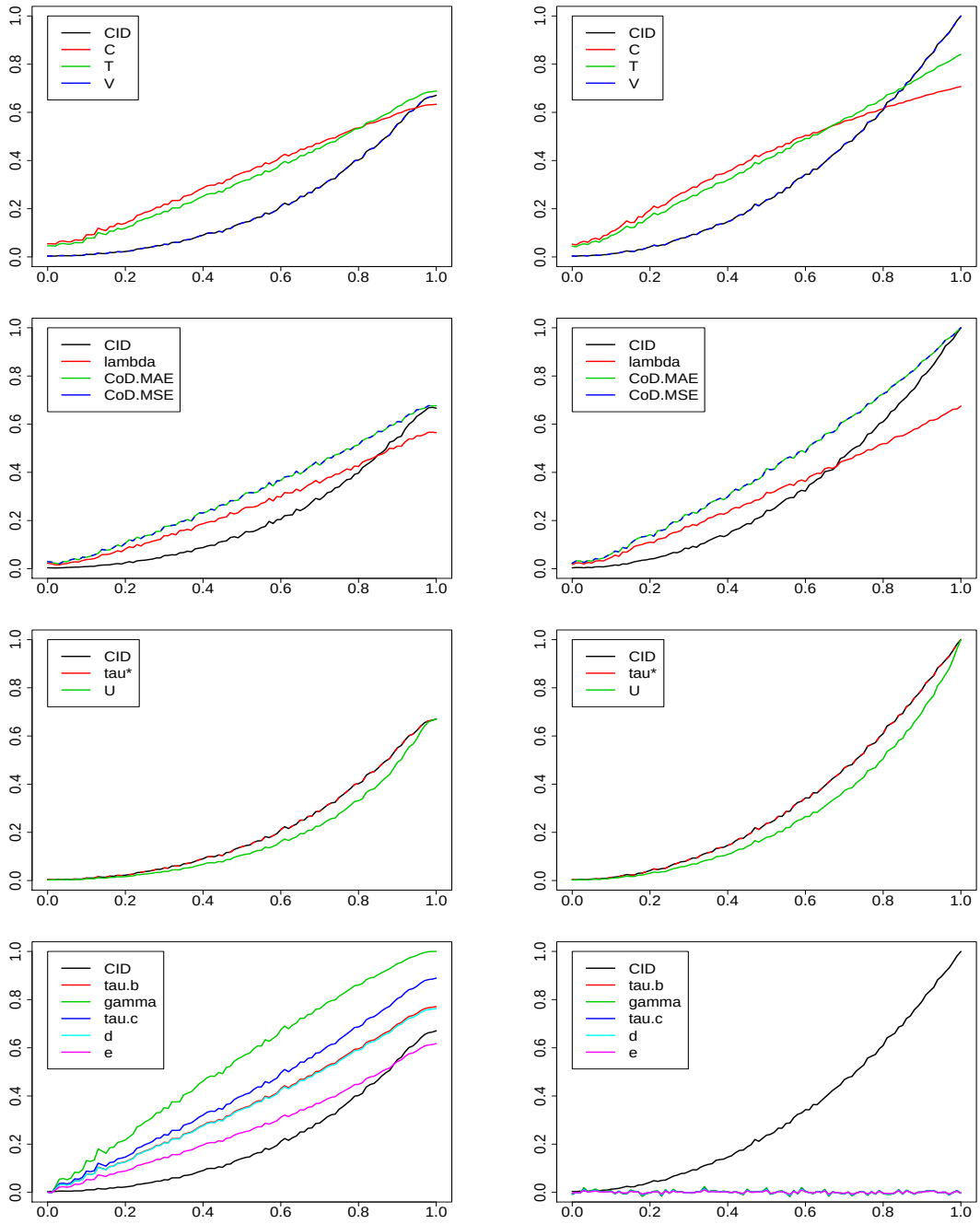


Figure 11: Four groups of association measures for two discrete variables are compared with CID. The left panel are the experimental results when the linear model (Design 1) is applied. The plots on the right panel illustrates the results of curvilinear model (Design 4). From top to bottom, the plots display the results for measures based on χ^2-statistics, measures based on optimal prediction, measures based on variance reduction, and measures for two ordinal variables. The data consists of a two-category target and a three-categorical predictor.

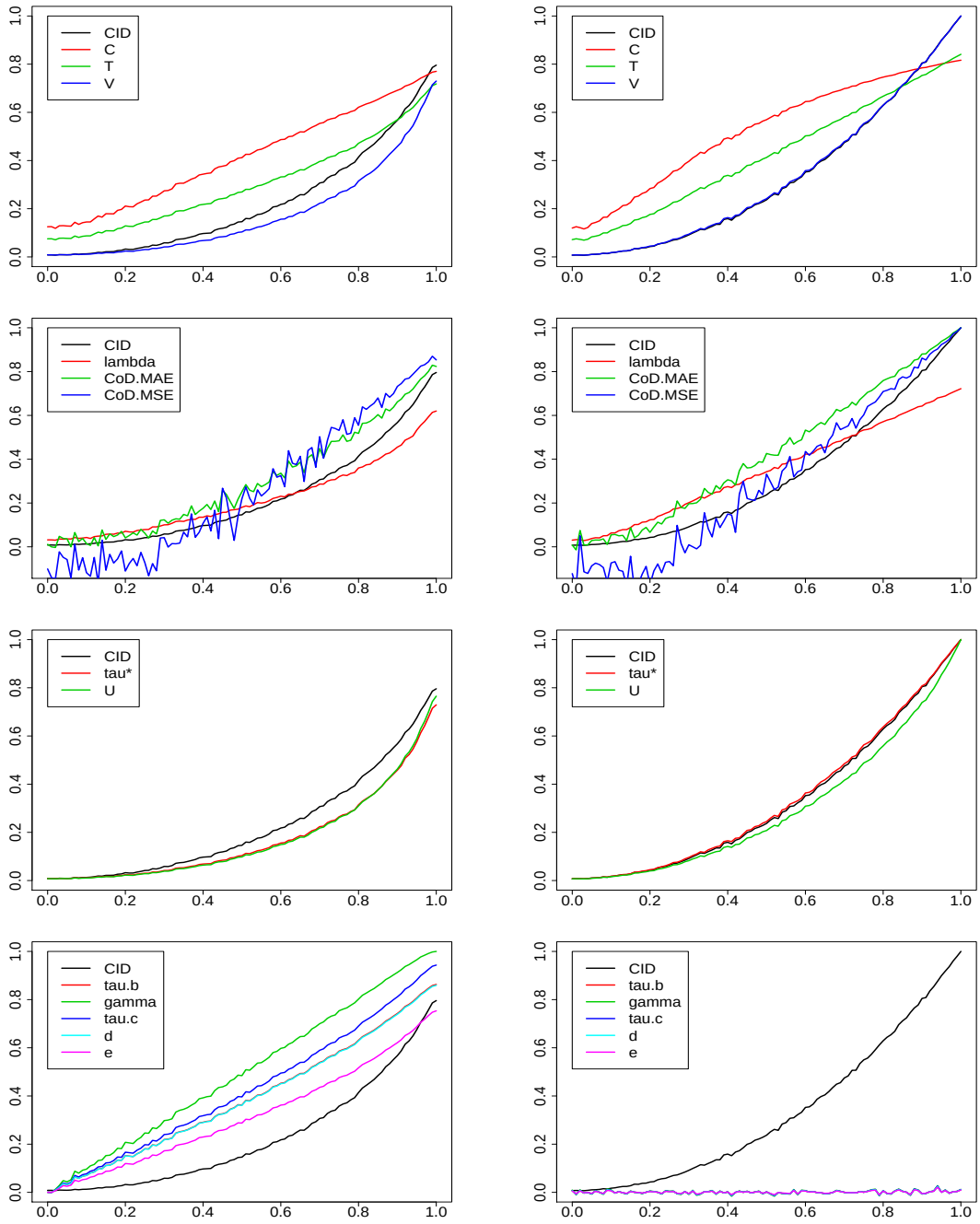


Figure 12: Four groups of association measures for two discrete variables are compared with CID. The left panel are the experimental results when the linear model (Design 1) is applied. The plots on the right panel illustrates the results of curvilinear model (Design 4). From top to bottom, the plots display the results for measures based on χ^2-statistics, measures based on optimal prediction, measures based on variance reduction, and measures for two ordinal variables. The data consists of a three-category target and a five-categorical predictor.

variable, one way to imitate the correlation is to account the ranks of observations for the computation of correlation (r_{DR} and r_{RI}). It also can be the preciseness of prediction of the ordering be considered (τ_b, τ_c, d, e, and γ). Other methods assign the averages of the other dichotomous or interval variable to the multichotomous variable for the computation of correlation (r_{MD}, r_{MR}, and η). Yet the monotonicity is not the only way to inspect the association among categorical variables. Other measures of association take advantage of the fact that a discrete variable can be categorized into a finite number of subgroups and take into account of the goodness of fit (C, T, V), the accuracy of prediction (λ and CoD), or the reduction of variance (τ^* and U).

There are three restrictions that ought to be considered when applying the above measures: (1) the normality assumption, (2) the linearity or monotonicity, and (3) the data characteristic. People should be extremely cautious that the correlation coefficient are limited by all of the three conditions. The other correlation indices (Group I, II, and III in Table 4 except ρ) and the measures for ordinal variables (Group V in Table 4) devote themselves to discern the pattern of linearity or monotonicity in the sample of different data types and, undoubtedly, they will do a wonderful job when the truth is corresponding to a linear model. However, the simulation results show that their strength is also their weakness: they have difficulties to recognize a nonlinear or non-monotone activities. Measures in Group IV in Table 4 are much more flexible in the sense of determining the existence of relationship. Although one restriction appears in the application of those measure in Group IV: they are only suitable for categorical data. Surely people can artificially categorize a continuous variable but the loss of information may be lethal. CID is free of the above restrictions. It requires no distributional or functional assumptions and can be easily implemented to either continuous or categorical data. More importantly, CID can be extended to multivariate analysis while the traditional measures cannot.

People also require an association measure having the ability to differentiate different levels of dependence. It helps people to make decisions about choosing essential predictors from all of the candidates. In a curvilinear relationship, it is shown by the simulation results that CID is definitely competent to deal with differential matters. While many traditional association measures constantly yield the value 0 regardless of the level of dependence, CID nondecreasingly goes from 0 to 1. Another nice property of CID can be observed via simulations: the aggravation of CID starts gradually and then becomes more rapid when the dependence is more severe. This can be beneficial to separate highly associated predictors from mildly associated predictors or even rank those highly associated predictors. It usually is the highly associated predictor of people's interest since the information from mildly associated predictors do not improve the understanding of the target variable as much as highly associated predictors do. It motivates the application of CID on variable selection. More details are followed in the next chapter.

CHAPTER IV

VARIABLE SELECTION

4.1 Introduction

In some studies, especially observational ones, researchers may measure a large number of potential predictors in their attempt to include all relevant ones. A natural question is therefore raised if keeping only a few of the predictors will not reduce the ability of making further inference. Retaining predictors that are not contributing is undesirable, especially when they are difficult or costly to measure. Besides, it is easier to handle a smaller number of predictors. Reducing variables also makes the model more parsimonious. One example is the analysis of microarray data (Duggan et al., 1999; Schena et al., 1995). Even though a massive number of genes are inspected at the same time, recent studies suggest that a few (2 or 3) genes have sufficient information (Li and Yang, 2000; Xiong et al., 2001).

Suppose we have measured m predictors and there is a suitable function g of k of these predictors so that

$$Y = g(X_1, X_2, \cdots, X_k) + \epsilon,$$

where ϵ is the measurement error. Our task is to find the function g and to identify the set of k relevant variables. It is sometimes referred as a *prediction* problem if the target variable is continuous and as a *classification* problem if the target variable is categorical. Usually, people begin with the search of a proper function g and determine the essential variables later according to a certain criterion. In the application of linear models, the search for g is restricted to the class of linear models. A number of criteria, such as R^2, adjusted R^2, or Mallows' C_p, help to identify the most rele-

vant variables while some popular sweep methods help to evaluate the combinations of predictors (e.g., Hocking, 1996). Variable selection also appears in the context of feature selection in pattern recognition. People wish to select a subset of predictors that provides an optimal classifier with minimum misclassification rate for the target variable. In such cases, the function g (or "classifier", as is called in the feature selection literature) can be selected from a more wide-ranging class. Given the classifier, the features which produce the smallest classification error are collected. Moreover, different choice of models greatly affects the results (Jain and Zongker, 1997).

Another intuitive solution is to select a variable subset of size k at first by considering the magnitude of dependence between the target variable and all possible combinations of k predictors. The best one or few subsets of size k with the highest level of dependence are potential candidates to put an interpretation on the target variable. Once the k most relevant predictors are identified, a variety of parametric or nonparametric approaches can be used to make further inference about the true model g. In that regard, we propose to use CID as a association measure to compare different possibilities of the true model. The estimate of CID can be computed from the data. The larger value of the CID estimate the stronger the association. However, large numbers of predictors are usually involved in variable selection problems. Due to the "curse of dimensionality", i.e. the sparsity of the data in a high dimensional space, it takes a huge sample to well estimate the true CID. Fortunately, estimating the true CID is almost never the goal but comparing the strengths of dependence is. In this chapter, we will demonstrate CID's capability to select variables in both simulations and the analysis of data from a study of breast cancer gene expression (van de Vijver et al., 2002).

4.2 Variable Selection for Classification

Consider the model

$$Y = I(cX * W + (1 - c)W * Z > 0), \tag{4.1}$$

where X, W, Z are independent samples of size 100 from the standard normal and $c \in [0, 1]$. Clearly,

$$\text{CID}(Y|X, W) > \text{CID}(Y|W, Z) \text{ iff } c > 1/2 \tag{4.2}$$

$$\text{and} \quad \text{CID}(Y|X, Z) < \max(\text{CID}(Y|X, W), \text{CID}(Y|W, Z)). \tag{4.3}$$

To determine the most influential predictor subset of variables, the linear discriminant analysis (LDA) is one of the most widely used method (Breiman et al., 1984). It assumes that all objects are from normal distributions with a common covariance matrix and determines the best classifier for each subset of predictors. The subset that produces the least classification error is chosen. After determining the classifiers, resubstitution or cross-validation can be used to estimate misclassification rates. However, LDA is problematic in this example due to the nonlinear structure in the model. Figure 13 shows the probabilities that the the pair of the most influential subset is correctly identified by LDA in 1000 simulations with 100 observations. The chances of selecting the correct subset according to either method are always less than 0.5 no matter what the value c is.

Alternatively, one could first identify essential predictors without assuming any specific model to the data by estimates of CID. We want to illustrate that, with high probabilities, we are able to detect the most influential subset(s) of predictors even though only a limited number of observations are available. The estimated CID profile curves given any two of the three predictors with 3 bins are presented in

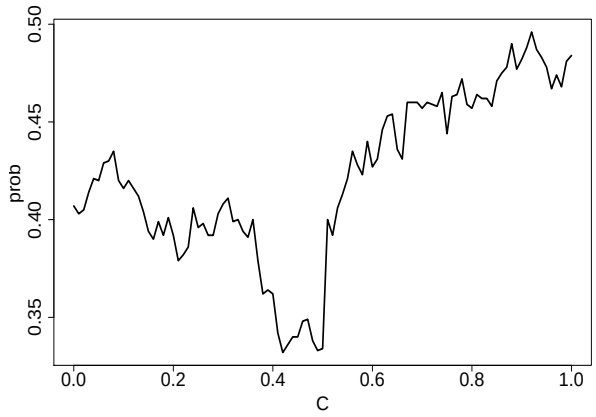


Figure 13: The plot describes the proportion that the correct pair of predictors had been selected based on LDA in 1000 simulations of Model (4.1) with different number of c while sample size is 100. The correct pair of predictors has to satisfy both of (4.2) and (4.3).

Figure 14. The true CID's from bootstrapping are indicated in dashed lines. It is not a surprise that the estimated CID's are not that close to the truth with such small sample size. However, the estimated CID of the most influential pair tend to be the largest among all estimated CID's Figure 15 displays the probability that the the pair of the most influential variables is correctly identified based on 1000 estimated CID with 100 observations and 3 bins. Observe that the chance of successful identification by CID is greater than 90 percent if $|c - 0.5| > 0.1$; the probability is getting higher while c is further from 0.5.

4.3 Variable Selection for Prediction

An example appeared in Friedman and Stuetzle (1981) is adopted here to demonstrate variable selection by CID in a prediction problem. Let

$$Y = X_1 X_2 + \epsilon \tag{4.4}$$

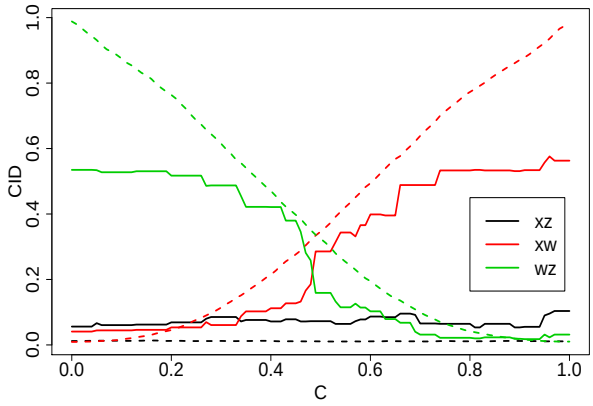


Figure 14: The solid curves in the plot are the estimated CID's of Y given two of the three predictors in Model (4.1) with sample size 100 and bin size 3 from only one simulation. CID$(Y|W,Z)$ (green) and CID$(Y|X,W)$ (red) are expected to stand out when c is less than 0.5 and greater than .5, respectively. The dashed curves indicate the true CID's obtained from bootstrapping.

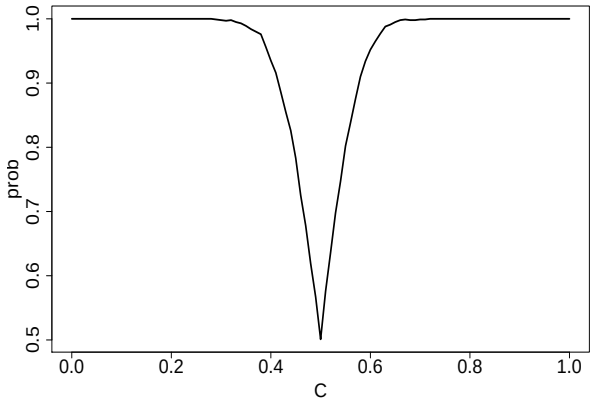


Figure 15: The plot describes the proportion that the correct pair of predictors had been selected based on estimated CID in 1000 simulations of Model (4.1) with different number of c while sample size is 100 and bin size is 3. The correct pair of predictors has to satisfy both of (4.2) and (4.3).

with (X_1, X_2) uniformly distributed on $(-1, 1) \times (-1, 1)$ and $\epsilon \sim N(0, 0.04)$. The original intentions in Friedman and Stuetzle (1981) were to identify a set of linear combinations of predictors and to model the regression surface as a sum of general smooth functions of the selected linear combinations. Their approach (named *projection pursuit regression*) simply performed nonparametric regression on the target given one of many possible linear combinations of the predictors. Whenever the smoothing function produced an error less than a pre-assigned threshold, the corresponding linear combination was selected. As a result, two linear combinations, $X_1 + X_2$ and $X_1 - X_2$, were selected since

$$X_1 X_2 = \frac{1}{4}(X_1 + X_2)^2 - \frac{1}{4}(X_1 - X_2)^2.$$

To ensure selecting all informative linear predictors, the researcher might put a lot of effort on numerous smoothing processes even though the smoothing results of irrelevant ones might be trashed in the end. Hence we ask ourself whether CID can assemble significant linear predictors without fitting models. Two simulations are conducted, one of which has the same setting as that in Friedman and Stuetzle (1981), whereas the other receives more noise by letting ϵ in Model (4.4) be taken from $N(0, 1)$. In each simulation, a sample of size 1000 is generated. We compute estimates of $\mathrm{CID}(Y|X_1 + rX_2)$ with bin size 20 for all r between -4 and 4. The results of two experiments are displayed in Figures 16 and 17, respectively. Both plots visualize that $X_1 + X_2$ and $X_1 - X_2$ are the most informative combinations. More importantly, even one has a noisy data as the second simulation does, CID can still provide some clues about which range of r's to look into.

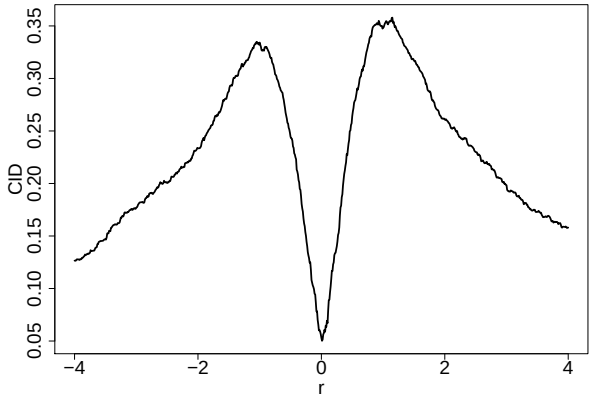


Figure 16: One sample of size 1000 from Model (4.4) is generated. The plot displays all $\text{CID}(Y|X_1 + rX_2)$ estimates for $r \in [-4, 4]$ in order to search the best linear combinations of X_1 and X_2 to be used in a smooth function to model the target variable Y.

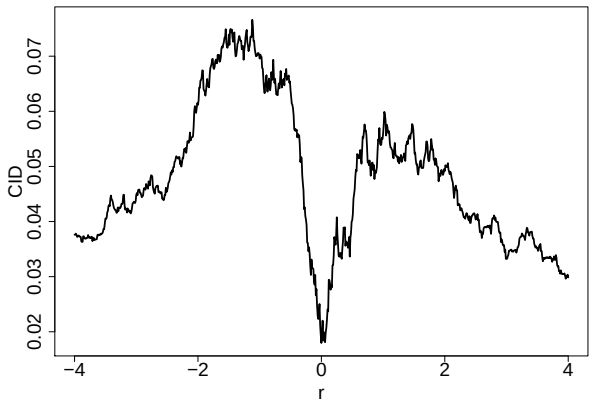


Figure 17: One sample of size 1000 from Model (4.4) is generated except ϵ is now from $N(0, 1)$. The plot displays all $\text{CID}(Y|X_1 + rX_2)$ estimates for $r \in [-4, 4]$ in order to search the best linear combinations of X_1 and X_2 to be used in a smooth function to model the target variable Y.

4.4 Genomic Application

Here we apply CID in two contexts: (a) expression-based classification, and (b) measurement of relationships among genes via prediction. Researchers have paid much attention to these problems especially after the microarray technology which can simultaneously monitor the mRNA expression levels of an enormous number of genes have been developed (Mukherjee et al., 2003). Classification tools are useful in discriminating the morphology of a sample and making a diagnosis, especially in clinical oncology (e.g., Beer et al., 2002; Huang et al., 2003; Gordon et al., 2002; Kari et al., 2003; Nutt et al., 2003; Shipp et al., 2002; van't Veer et al., 2002; Wigle et al., 2002). Also, the understanding of inter-gene relationships might be beneficial to building up the genetic regulatory networks (Holter et al., 2001). Both of these problems point to the same issue: measuring multivariate association among variables, and this is what the CID is designed for.

In this section, we demonstrate the analysis by using a data set taken from a study of breast cancer (van de Vijver et al., 2002). This research attempted to determine a more powerful predictor of the disease based on gene expression levels than standard systems based on clinical and histological criteria. The data consists of the 295 breast-cancer patients who were 53 years old or younger. From each patient, RNA was isolated from tumor material and used to derive complementary RNA (cRNA). A fluorescent dye reversal technique was applied to the cRNA microarrays and the fluorescence intensities of scanned images were qualified, normalized and corrected. There were approximately 25,000 genes being monitored for each patient.

In a previous study (van't Veer et al., 2002), about 5,000 genes which were significantly regulated were selected from the 25,000 or so genes on the microarray. The correlation coefficient of the expression for each gene with disease outcome

was calculated and 231 genes were found to be significantly associated with disease outcome (correlation coefficient < -0.3 or > 0.3). These 231 genes were then rank-ordered on the basis of the magnitude of the correlation coefficient. The top 70 genes in the rank-ordered list were then found to be the most powerful set of genes to correctly classify prognosis.

There are 234 of 295 profiles were newly developed by van de Vijver et al. (2002). They calculated the correlation coefficient of the level of expression of the 70 genes with the average profile of these genes in tumors from patients with a good prognosis in van't Veer et al. (2002). A patient with a correlation coefficient of more than 0.4 was then assigned to the group with a good-prognosis signature, and all other patients were assigned to the group with a poor-prognosis signature. The other 61 profiles were extracted from the study conducted by van't Veer et al. (2002). For these 61 profiles, they used a threshold of 0.55 of correlation coefficient to determine which prognosis group the patient was assigned to. As a result of classification, there are 115 patients having good-prognosis signature and 180 patients having poor-prognosis signature. We have access only to the gene expression data of the 70 genes and therefore we will limit our analysis to this.

4.4.1 Classification

In the application of classification, we intend to find good classifiers of prognosis signature of size 2 based on CID. A dummy variable, Y, is created to specify the patients' prognosis signature: "0" indices a good one and "1" indices a bad one. Letting the binary Y as the target variable, we compute the CID estimates given the log-intensities of any of $\binom{70}{2}$ paired predictors with 3 bins each. The predictor set is determined to be the best to classify the prognosis signature if it has the highest CID estimates; the second best predictor set has the the second highest CID estimates and

so on. Among those 2415 paired predictors, the combination of Genes 49 (ORC6L) and 60 (IGFBP5) becomes the best classifier of size 2 with the highest CID estimate (0.5104). In Figure 18, we plot the log-intensity of Gene 60 (IGFBP5) against that of Gene 49 (ORC6L). It shows that the patients of good or bad prognosis have been well separated. In the other hand, the set of Genes 17 (FLT1) and 67 (SM-20) has the lowest CID estimate (0.0107) and hence is declared to be the worst classifier for prognosis signature. Figure 19 visualizes the combinations of Genes 17 (FLT1) and 67 (SM-20) lacks power to differentiate the patients of good or bad prognosis.

To make a sense about how well CID does, we obtain the mean absolute errors (MAE) for each classifier by "leave-one-out" method from 3-nearest-neighbor classification. The MAE's have been plotted against the estimates of CID in Figure 20. Apparently, a high estimate of CID corresponds to a low MAE and vice versa. For instance, the best two-gene classifier of prognosis signature, Genes 60 and 49, has MAE equal to 0.1390, which is the second lowest MAE among those for all two-gene classifiers. The MAE of the worst two-gene classifier according to CID, Genes 17 and 67, is 0.4814 and is one the highest MAE's.

4.4.2 Prediction

The CID can be used to catch primitive explanatory variables in a prediction problem. Here we try to explore the relationship among expression of genes. We use expression data from the aforementioned breast-cancer microarray study involving 295 patients for predicting the expression level of a target gene via the level of two predictor genes. For each of the 70 genes, $\binom{69}{2}$ CID estimates given two of the remaining 69 genes using the log intensities are calculated with 3 bins each. We observe that letting Gene 66 (CENPA) as the target variable yields the largest mean (0.3232) of CID estimates. Hence Gene 66 is selected for our demonstration. Among all 2346 two-gene sets to

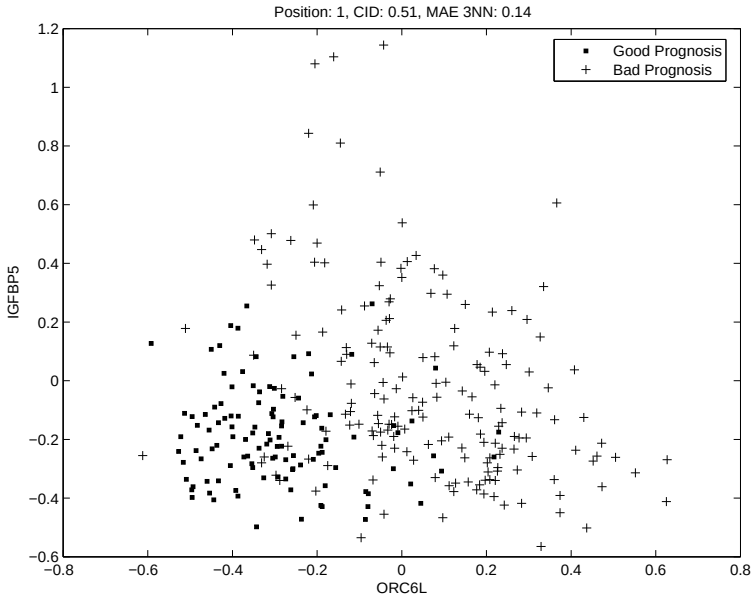


Figure 18: The plot shows the values of the best (Genes 49 and 60) set of predictors for the 295 patients and their associated prognosis signatures.

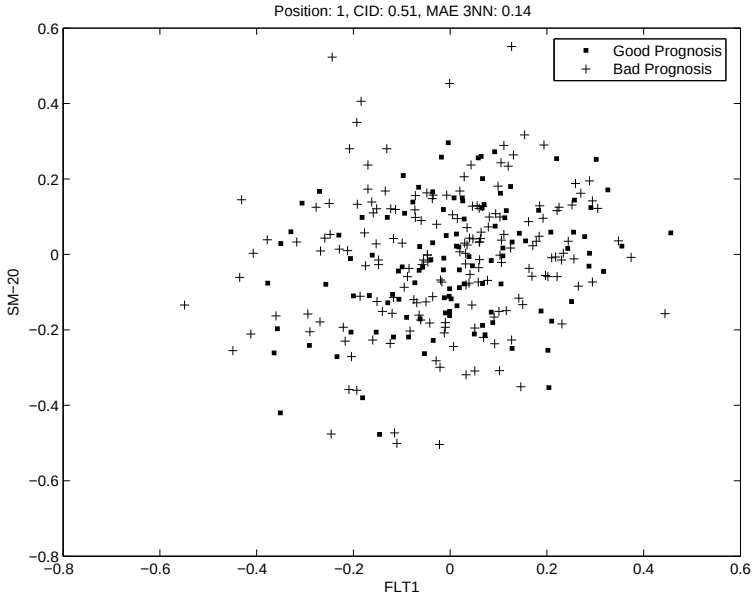


Figure 19: The plot shows the values of the worst (Genes 17 and 67) set of predictors for the 295 patients and their associated prognoses signatures.

predict Gene 66 (CENPA), the combination of Gene 40 (KIAA01775) and Gene 64 (PRC1) produces the highest CID estimate (0.5082) and the combination of Genes 57 (IGFBP5a) and 60 (IGFBP5b) produces the lowest (0.0145). They are claimed by CID to be the best and the worst predictor sets, respectively. Figure 21 presents the surface of predicting Gene 66 via the predictor gene set (40, 64) by neural network. It shows a good fit with the data points. The surface for the set (57, 60) is provided in Figure 22 as well. It agrees with CID that gene set (57, 60) is not a good choice of predicting Gene 66.

From a 2-layer neural network with three neurons in the inner layer, we estimate the mean squared error (MSE) for the prediction of Gene 66 given any predictor set of size 2. Note that a small number of neurons is used to avoid over-fitting. All 295 objects are used for design. The MSE is estimated by using resubstitution, which is close to unbiased with very small variance for 295 training examples. Figure 23 shows the scatter plot between CID estimate and the MSE computed from neural network. A fairly tight linear relationship presented in the scatter plot explains the strong agreement between the findings of CID and those of neural network. Particularly, the best two-gene predictor set found by CID, Genes 40 and 64, is corresponding to the lowest MSE (0.0246) while the MSE for the worst predictor set, Genes 57 and 60, is almost the highest (0.0993).

4.5 Discussion

This chapter has illustrated the application of CID on variable selection. There are two issues related to the variable selection: the prediction and classification problem. The CID has been firstly practiced in the simulations regarding to both issues. As noted before, estimation of CID cannot be achieved with samples that are small relative to the number of variables. However, estimation seldom is the objective of

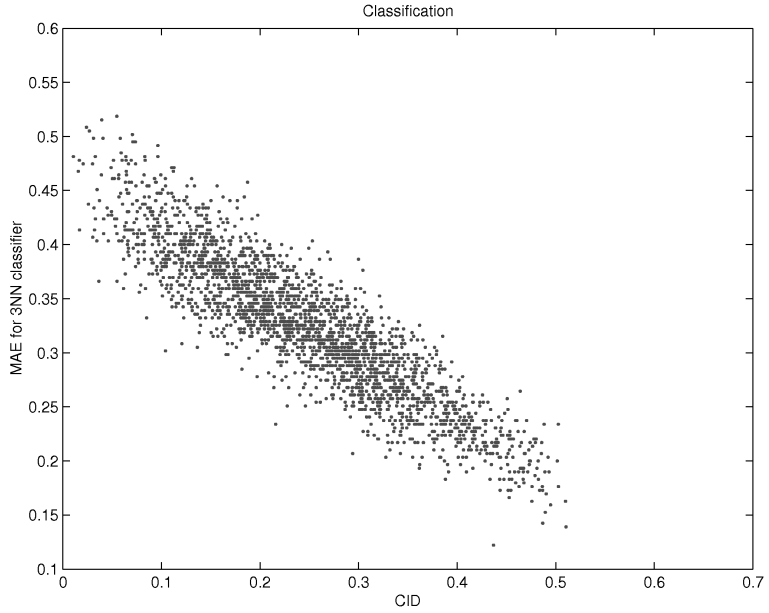


Figure 20: The scatter plot between the CID and the MAE from 3-nearest-neighbor classification for all two-gene classifiers.

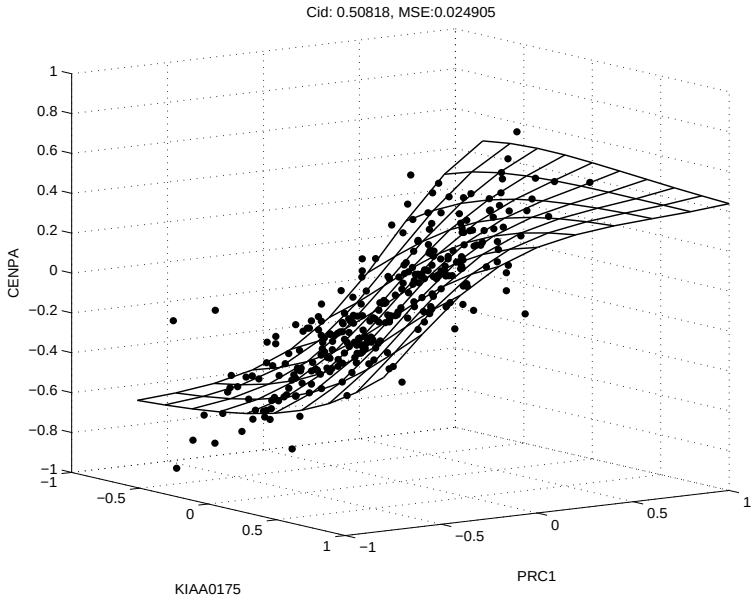


Figure 21: The plot shows the surface obtained by the neural network predicting Gene 66 via the best set of two predictors, Genes 40 and 64.

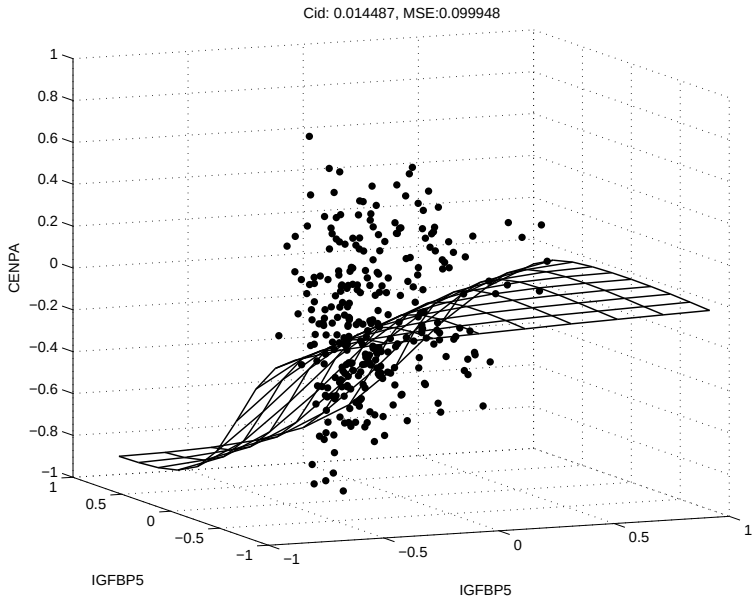


Figure 22: The plot shows the surface obtained by the neural network predicting Gene 66 via the worst set of two predictors, Genes 57 and 60.

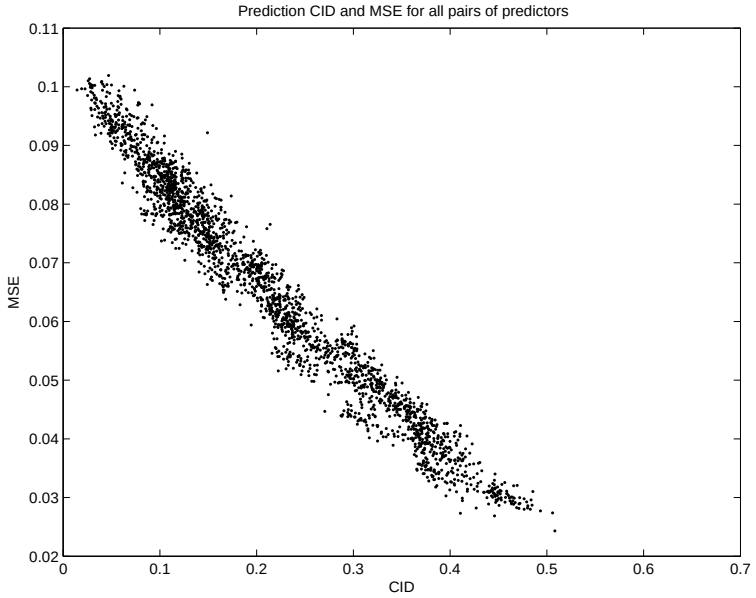


Figure 23: The scatter plot between the CID and the MSE estimated for a neural-network predictor of target Gene 66 for all predictor sets.

variable selection but to correctly identify essential features. The results of two simulations verify the capability of CID to recognize influential elements in the prediction and classification problem even only with a small sample.

The CID is also applied to two contexts of microarray studies: expression-based classification, and measurement of relationships among genes via prediction. We intend to classify the binary prognosis signature by complying the expression of two out of 70 genes regarding to the former issue and to predict the expression level of one gene by two of the rest 69 genes regarding to the later. We are pleased by the fact that the breast cancer data used in here is large compared to those of most microarray studies, where small-sample estimation is a common impediment (Dougherty, 2001). It is also our belief that CID can handle small samples with success. Observing the prediction surfaces and the classification diagrams suggests appropriateness of the selected features in both contexts. The nearest-neighbor classification and neural network individually support the CID's finding from two different aspects. Therefore, it can be concluded that CID is proficient in microarray-based quantification of gene interaction.

Note that the breast cancer data we are allowed to access has come up "clean" per se — only the expression of 70 genes that are mostly highly correlated with disease outcome are reported. It is hard to eliminate the thought that possible complex associations might exist among those excluded genes. Hence it has become our wish to have the opportunity exploring a more complete dataset for the intrinsic dependence with CID in the future.

CHAPTER V

SUMMARY

The objective of this study is to develop a measure of association. It is motivated by the long-time inquires of dependence measures from all disciplines. By reviewing the literature, we summarized in Introduction that an appropriate measure of association should be assumption-free, should be capable of differentiating various dependence levels, should be easily implemented in different occasions, an should be helpful for the investigation of causality.

In response to these inquires, we propose the coefficient of intrinsic dependence, or CID in abbreviation. It follows ritual definition of independence by quantifying the squared discrepancy between marginal and conditional distribution of the target variable(s). The squared discrepancy is then integrated and normalized to retain the convention about an association measure: it takes values between 0 and 1 inclusive; it is zero under independence and 1 under absolute dependence. The value of CID gradually increases as the explanatory variables assert larger influence on the target and it is invariant under monotone transformation. Numerous simulations in this dissertation demonstrate that CID is ready to explore the relationships among continuous variables, categorical variables, or mixtures of both. It can be fully extended to multivariate cases by implanting appropriate multivariate distribution functions. The CID also takes causal relationship into account because of its asymmetric property.

Prior to CID, researchers have already been provided many options of association measures. Due to the absence of the measure which can be globally applied to both continuous and discrete data, people has split the search for the most adequate measure based on data types. For continuous data, the correlation coefficient

is the most commonly used method. People profit from its naturalness as a measure of dependence in the most familiar multivariate normal distribution and straightforwardness of calculation. Moreover, the inference about the correlation coefficient requires strong normality assumption and it cannot describe anything but the linear relationship. While more and more people suspect the occasional existence of non-normality or nonlinearity in their studies, they turn to rank-based measures, such as Spearman's ρ and Kendall's τ statistic. The simulation results (Figure 8), however, show that these rank-based measures maybe stretch the normality assumption but are still harmed by nonlinearity; in the simulations, the correlation coefficient, Spearman's ρ, and Kendall's τ all yield the value approximately 0 in the quadratic relationship.

There have been significant developments of association measures for categorical data. The values of a categorical variables can be divided into a finite number of subgroups so that the frequency of the occurrence in each subgroup, or category, are usually of consideration. Among others, the ordinal variables preserve the ordering of the labels of their categories and it is usually the level of monotonicity between two variables being measured. Therefore, it is not surprising that a nonmonotonicity trend in the simulations would bring down the measures particulary for ordinal variables (Figures 11 and 12). A similar conclusion can be made for the correlation indices for ordinal or dichotomous variables (Figures 9 and 10). In the other hand, the correlation indices involving one or two multichotomous variables, including r_{MD}, r_{MR} and η, remain their ability to differentiate association levels in different settings. One reasonable explanation is their way to imitate the correlation coefficient when their categories do not have metric meanings. The average of the other non-multichotomous observations in one category is manually assigned to be the score of that category and it is the correlation between the artificial scores and the non-multichotomous

observations is computed as the association between two original variables. Chen and Popovich (2002) point out this kind of association is equivalent to assess

$$\sqrt{\frac{SS_{\text{between}}}{SS_{\text{total}}}}$$

and is able to describe the curvilinear relationship between a multichotomous variable and a variable measured at interval or ordinal level. The simulation results in Figures 11 and 12 cooperate with the above statement. There are more measures of association looking from different angles rather than monotonicity. Simulation results in Figures 11 and 12 indicate their robustness for different models as well. Unfortunately, these promising measures are only available for categorical data analysis.

Regarding to the issue of hypothesis tests, the sampling distribution of CID under independence has been imitated by simulations in Section 2.5. Accordingly, the estimate of $100(1-\alpha)\%$ quantile can be obtained for a level α hypothesis test of independence. Experiments demonstrate the power of the tests based on CID in different occasions that may be appreciated in the exploratory study while the underlying distribution is not fully comprehended (Figures 6 and 7). With similar argument, CID is recommended for variable selections. Figures 14 through 17 show that, although it may be problematic to obtain an accurate estimate of CID when the sample size is too small (due to the "curse of dimensionality"), CID remains the ability to single out the set of primary predictors. Observe the experiments in Chapter 3, along with the raise of dependence level, the increment of CID is gentle at first and then swift when the relationship getting stronger. It even provides a greater opportunity of successfully differentiate all feature highly dependent on the target variable(s).

Finally we apply CID to the actual microarray data conducted by van de Vijver et al. (2002). The data set contains the expression data of 70 genes from 295 patients.

In the practice of classification problem, we identify two genes, ORC6L and LGFBP5, is the best combination to classify prognosis signature and the combination of FLT1 and SM-20 appears to be the worst. Regarding to the prediction problem, it is our intention to find the combination of two genes which most possibly control the expression of gene CENPA. We find the combination of gene KIAA01775 and PRC1 is the best and the combination of IGFBP5a and IGFBP5b is the worst in prediction of CENPA. The errors of classification and prediction based on each pair of predictors are seized by 3-nearest-neighbor classification and neural network, respectively. The scatter plots of CID values against the corresponding errors (Figures 23 and 20) suggest these two methods from different aspects back up our findings based on CID.

REFERENCES

Adryan, B. and Schuh, R. (2004). Gene-ontology-based clustering of gene expression data. *Bioinformatics* **20**, 2851–2852.

Agresti, A. (1990). *Categorical Data Analysis*. New York, NY: John Wiley & Sons, Inc.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 6745–6750.

Antonov, A. V., Tetko, I. V., Mader, M. T., Budczies, J., and Mewes, H. W. (2004). Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics* **20**, 644–652.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., and Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* **8**, 816–824.

Bergmann, S., Ihmels, J., and Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. *PLoS Biology* **2**, 85–93.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth, Inc.

Broët, P., Lewin, A., Richardson, S., Dalmasso, C., and Magdelenat, H. (2004). A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics* **20**, 2562–2571.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. J., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 262–267.

Brutlag, J. D. (1998). *History of Correlation and Association: The Development of Correlation and Association in Statistics*. Available at http://www.geocities.com/j_brutlag/corr.htm.

Carreira-Perpiñán, M. Á. (1997). *A Review of Dimension Reduction Techniques*. Technical Report CS-96-09, Department of Computer Science, University of Sheffield, U.K.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Belmont, CA: Wadsworth, Inc.

Chen, D.-T., Lin, S.-H., and Soong, S.-J. (2004). Gene selection for oligonucleotide array: an approach using PM probe level data. *Bioinformatics* **20**, 854–862.

Chen, P. Y. and Popovich, P. M. (2002). *Correlation: Parametric and Nonparametric Measures*. Sage University Papers Series on Quantitative Applications in Social Sciences, 07-139. Thousand Oaks, CA: Sage Publications, Inc.

Cheng, R. C. H. and Jones, O. D. (2004). Analysis of distributions in factorial experiments. *Statistica Sinica* **14**, 1085–1103.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-Fit Techniques.* New York, NY: Marcel Dekker, Inc.

Dougherty, E. R. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics* **2**, 28–34.

Dougherty, E. R., Kim, S., and Chen, Y. (2000). Coefficient of determination in nonlinear signal processing. *Signal Processing* **80**, 2219–2235.

Draper, N. and Smith, H. (1998). *Applied Regression Analysis,* 3rd edition. New York, NY: John Wiley & Sons, Inc.

Duggan, D. J., Bittner, M. L., Chen, Y., Meltzer, P. S., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics* **21**, 10–14.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

Ermolaeva, O., Rastogi, M., Pruitt, K. D., Schuler, G. D., Bittner, M. L., Chen, Y., Simon, R., Meltzer, O., Trent, J. M., and Boguski, M. S. (1998). Data management and analysis for gene expression arrays. *Nature Genetics* **20**, 19–23.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76**, 817–823.

Galton, F. (1889). Correlation and their measurement, chiefly from anthropometric data. *Nature* **39**, 238.

Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 12079–12084.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Goodman, L. A. and Kruskal, W. H. (1979). *Measures of Association for Cross Classfication*. New York, NY: Springer-Verlag Inc.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., and Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research* **62**, 4963–4967.

Granger, C. W., Maasoumi, E., and Racine, J. (2004). A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis* **25**, 649–669.

Greenland, S. (1996). A lower bound for the correlation of exponentiated bivariate normal pairs. *The American Statistician* **50**, 163–164.

Hocking, R. R. (1996). *Methods and Applications of Linear Models*. New York, NY: John Wiley & Sons, Inc.

Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V., and Banavar, J. R. (2001). Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 1693–1698.

Hsing, T., Liu, L.-Y., Brun, M., and Dougherty, E. R. (2005). The coefficient of intrinsic dependence (feature selection using el CID). *Pattern Recognition* **38**, 623–636.

Huang, E., Cheng, S. H., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iverson, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R., and Huang, A. T. (2003). Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596.

Jain, A. and Zongker, D. (1997). Feature selection: Evaluation, application and small sample performance. *IEEE Transactions on Pattern Anaylysis and Machine Intelligence* **19**, 153–158.

James, L. R., Mulaik, S. A., and Brett, J. M. (1982). *Causal Analysis: Assumptions, Models and Data*. Beverly Hills, CA: Sage Publications, Inc.

Kari, L., Loboda, A., Nebozhyn, M., Rock, A. H., Vonderheid, E. C., Nichols, C., Virok, D., Chang, C., W, H, H., Johnson, J., Wysocka, M., Showe, M. K., and Showe, L. C. (2003). Classification and prediction of survival in patients with the leukemic phase of cutaneous T cell lymphoma. *Journal of Experimental Medicine* **197**, 1477–1488.

Kat, H. M. (2003). The dangers of using correlation to measure dependence. *Journal of Alternative Investments* **6**, 54–58.

Li, W. and Yang, Y. (2000). How many genes are needed for a discriminant microarray data analysis. In *Proceedings of the First Conference on Critical Assessment of Microarray Data Analysis*. CAMDA 2000.

Liebetrau, A. M. (1983). *Measures of Association*. Sage University Papers Series on Quantitative Applications in Social Sciences, 07-001. Beverly Hills, CA: Sage Publications, Inc.

Macdonell, W. R. (1902). On the influence of previous vaccination in cases of small-pox. *Biometrika* **1**, 375–383.

Mari, D. D. and Kotz, S. (2001). *Correlation and Dependence*. London: Imperial College Press.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* **105**, 156–166.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology* **10**, 119–142.

Neuhäuser, M. and Lam, F. C. (2004). Nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. In *Proceedings of the 2nd Asia-Pacific Bioinformatics Conference*, Vol. 29, 139-143. Dunedin, New Zealand: Australian Computer Society.

Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., Black, P. M., von Deimling, A., Pomeroy, S. L., Golub, T. R., and Louis, D. N. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research* **63**, 1602–1607.

Pedhazur, E. J. and Schmelkin, L. P. (1991). *Measurement, Design, and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum.

Persons, W. M. (1916). The construction of a business barometer based upon annual data. *American Economic Review* **6**, 739–769.

Rasmussen, J. L. (1986). An evaluation of parametric and non-parametric tests on modified and non-modified data. *British Journal of Mathematical and Statistical Psychology* **39**, 213–220.

Richardson, M. W. and Stalnaker, J. M. (1933). A note on the use of biserial r in test research. *Journal of General Psychology* **8**, 463–465.

Rodgers, J. L. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician* **42**, 59–66.

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* **270**, 467–470.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., and Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* **8**, 68–74.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* **15**, 72–101.

Stigler, S. M. (1973). Simon newcomb, percy daniell, and the history of robust estimation. *Journal of the American Statistical Association* **68**, 872–879.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. London: Belknap Press.

Szabo, A., Perou, C. M., Karaca, M., Perreard, L., Quackenbush, J. F., and Bernard, P. S. (2004). Statistical modeling for selecting housekeeper genes. *Genome Biology* **5**, R59.

Thomas, J. G., Olson, J. M., Tapscott, S. J., and Zhao, L. P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* **11**, 1227–1236.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Astma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347**, 1999–2009.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van de Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bemards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

Wherry, R. J. (1984). *Contributions to Correlational Analysis*. London: Academic Press, Inc.

Wigle, D. A., Jurisica, I., Radulovich, N., Pintilie, M., Rossant, J., Liu, N., Lu, C., Woodgett, J., Seiden, I., Johnston, M., Keshavjee, S., Darling, G., Winton, T., Breitkreutz, B.-J., Jorgenson, P., Tyers, M., Shepherd, F. A., and Tsao, M. S. (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research* **62**, 3005–3008.

Wilcox, R. R. (1990). Comparing variances and means when distributions have non-identical shapes. *Communications in Statistics – Simulation and Computation* **19**, 155–173.

Xiong, M., Fan, Z., and Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Research* **11**, 1878–1887.

Zhang, M. Q. (1999). Large-scale gene expression data analysis: A new challenge to computational biologists. *Genome Research* **9**, 681–688.

APPENDIX A

SOME DERIVATIONS OF CID

A.1 Derivation of Example 2.1

The details are included in Appendix in Hsing et al. (2005).

A.2 Derivation of Example 2.2

Let

$$Y|X \sim exp(x) \quad \text{and} \quad X \sim gamma(\alpha, \beta).$$

Denote the conditional distribution of Y given X by

$$f_{Y|X}(y) = x \exp\{-xy\}, \quad 0 < y < \infty.$$

The marginal pdf, $f_Y(y)$, and cdf, $F_Y(y)$, of Y are

$$\begin{aligned} f_Y(y) &= \alpha\beta(1+y\beta)^{-\alpha-1}, \quad 0 < y < \infty, \\ \text{and} \quad F_Y(y) &= 1-(1+y\beta)^{-\alpha} \end{aligned}$$

Therefore, the inverse function of cdf of Y can be written as

$$F_Y^{-1}(u) = \frac{1}{\beta}\left[(1-u)^{-1/\alpha} - 1\right].$$

Now

$$\begin{aligned} P(Y \le F_Y^{-1}(u)|X) &= 1 - \exp\{-x[(1-u)^{-1/\alpha} - 1]/\beta\}; \\ P^2(Y \le F_Y^{-1}(u)|X) &= 1 - 2\exp\{-x[(1-u)^{-1/\alpha} - 1]/\beta\} \\ &\quad + \exp\{-2x[(1-u)^{-1/\alpha} - 1]/\beta\}. \end{aligned} \tag{A.1}$$

From Equation (A.1), it can be easily shown that

$$\begin{aligned} \mathrm{E}[P^2(Y \le F_Y^{-1}(u)|X)] &= 2u - 1 + [2(1-u)^{-1/\alpha} - 1]^{-\alpha}, \\ \text{and} \quad \int_0^1 \mathrm{E}[P^2(Y \le F_Y^{-1}(u)|X)]du &= \int_0^1 [2(1-u)^{-1/\alpha} - 1]^{-\alpha} du. \end{aligned}$$

Hence the numerator of CID is

$$\int_0^1 [2(1-u)^{-1/\alpha} - 1]^{-\alpha} du - 1/3;$$

$$\mathrm{CID}(Y|X) = 6 \int_0^1 [2(1-u)^{-1/\alpha} - 1]^{-\alpha} du - 2.$$

A.3 Derivation of Example 2.3

In Example 2.3, we consider the cases of binary target variables. Without loss of generality, assume that the two possible values of the target variable, Y, are 1 and 2. Let

$$P(Y = 1) = p \quad \text{and} \quad P(Y = 1|X) = p_x.$$

Then the denominator of CID is

$$p^2(1-p) = pV(Y).$$

Also

$$\sum_{j=1}^{2} p_j \mathrm{E}[q_j^+(X)(1 - q_j^+(X))] = p\mathrm{E}[p_x(1-p_x)] = p\mathrm{E}[V(Y|X)].$$

Therefore,

$$\mathrm{CID}(Y|X) = 1 - \frac{p\mathrm{E}[V(Y|X)]}{pV(Y)} = 1 - \frac{\mathrm{E}[V(Y|X)]}{V(Y)} = \frac{V[\mathrm{E}(Y|X)]}{V(Y)}. \tag{A.2}$$

It is a well known result that if X is also binary then $\mathrm{CID}(Y|X) = r^2$, where r is the correlation coefficient. If $X \sim beta(a, b)$ and $Y|X \sim Bernoulli(x)$ then

$$V(Y) = \frac{ab}{(a+b)^2}; \quad V[\mathrm{E}(Y|X)] = \frac{ab}{(a+b)^2(a+b+1)}.$$

So that

$$\mathrm{CID}(Y|X) = \frac{1}{a+b+1}.$$

A.4 Derivation of Example 2.4

The result will be illustrated in a form of matrices. Let $\mathbf{p}$ and $\mathbf{p}_i$ denote the marginal and conditional mass functions of Y, respectively:

$$\begin{aligned}
\mathbf{p} &= (0.18 \quad 0.32 \quad 0.50)^T \\
\mathbf{p_1} &= (0.09 \quad 0.25 \quad 0.38)^T \ /0.72 \\
\mathbf{p_2} &= (0.09 \quad 0.07 \quad 0.12)^T \ /0.28
\end{aligned}$$

Suppose $\mathbf{1}_n$ is a $n \times 1$ vector of ones and $\mathbf{L}_n$ is a $n \times n$ matrix having the value one in its lower triangle and diagonal and the value zero in its upper triangle. The denominator of CID is

$$\begin{aligned}
(\mathbf{L}_3\mathbf{p})^T diag(\mathbf{p})(\mathbf{1}_3 - \mathbf{L}_3\mathbf{p}) &= (0.18 \quad 0.50 \quad 1.00) \begin{pmatrix} 0.18 & 0 & 0 \\ 0 & 0.32 & 0 \\ 0 & 0 & 0.50 \end{pmatrix} \begin{pmatrix} 0.82 \\ 0.50 \\ 0.00 \end{pmatrix} \\
&= 0.1066.
\end{aligned}$$

The numerator of CID is

$$0.1066 - 0.72 \times (\mathbf{L}_3\mathbf{p}_1)^T diag(\mathbf{p})(\mathbf{1}_3 - \mathbf{L}_3\mathbf{p}_1) - 0.28 \times (\mathbf{L}_3\mathbf{p}_2)^T diag(\mathbf{p})(\mathbf{1}_3 - \mathbf{L}_3\mathbf{p}_2)$$

$$= 0.1066 - 0.72 \times 0.0994 - 0.28 \times 0.1176 = 0.0020$$

Note that $\mathbf{L}_3\mathbf{p}$ and $\mathbf{L}_3\mathbf{p}_i$ can be considered as the corresponding cumulative mass functions of Y. Therefore,

$$\mathrm{CID}(Y|X) = 0.0020/0.1066 = 0.019.$$

A similar argument calculates $\mathrm{CID}(X|Y)$ while Equation (A.2) in Appendix A simplifies the computation for binary X.

APPENDIX B

R PROGRAMS

```
dCvM<-function(x,y,mx){ # m=number of bins
                        # y is a vector but x can be a nxp matrix
   z<-sort(unique(y))
   nz<-length(z)
   n<-length(y)

   # create the labels
   idx0<-c(1:(mx-1))/mx

   if (length(c(x))>n){ # when X is a matrix
      p<-ncol(x)
      xx<-x*0
      for (i in 1:p){
         xx[,i]<-rank(x[,i])
      }
      xx<-xx/n

      idxall<-x*0
      idx<-(1:n)*0
      for (i in 1:n){
         for (j in 1:p){
            idxall[i,j]<-length(idx0[idx0<xx[i,j]])+1
         }
      }
      uniq.x<-unique.array(idxall)
      ncx<-nrow(uniq.x)
      for (i in 1:ncx){
         x0<-uniq.x[i,]
         tt<-c(abs(idxall-t(matrix(x0,p,n)))%*%matrix(1,p,1))
         idx[tt==0]<-i
      }
   }
   else { # when x is a vector
      xx<-rank(x)/n
      idxall<-idx<-(1:n)*0
      for (i in 1:n){
         idxall[i]<-length(idx0[idx0<xx[i]])+1
      }
      uniq.x<-unique(idxall)
      ncx<-length(uniq.x)
      for (i in 1:ncx){
```

```
      x0<-uniq.x[i]
      idx[idxall==x0]<-i
   }
}


L<-lower.tri(matrix(0,length(z),length(z)))+diag(length(z))

# denominator
py<-tabulate(match(y,z),nb=nz)/n
Lpy<-L%*%py
den<-t(Lpy)%*%diag(c(py))%*%(1-Lpy)

# numerator
cx<-tabulate(idx)/n
vareg<-0
for (i in 1:ncx){
   sy<-y[idx==i]
   spy<-tabulate(match(sy,z),nb=nz)/length(sy)
   Lspy<-L%*%spy
   vareg<-vareg+(t(Lspy)%*%diag(c(py))%*%(1-Lspy))*cx[i]
}
num<-vareg


# integrate
out<-1-num/den
return(out)
}


dCvMbin<-function(x,y,mx){ # m=number of bins
                      # y is a vector but x can be a nxp matrix
z<-unique(y)
n<-length(y)

# create the labels
idx0<-c(1:(mx-1))/mx

if (length(c(x))>n){ # when X is a matrix
   p<-ncol(x)
   xx<-x*0
   for (i in 1:p){
      xx[,i]<-rank(x[,i])
   }
   xx<-xx/n
```

```
      idxall<-x*0
      idx<-(1:n)*0
      for (i in 1:n){
         for (j in 1:p){
            idxall[i,j]<-length(idx0[idx0<xx[i,j]])+1
         }
      }
      uniq.x<-unique.array(idxall)
      ncx<-nrow(uniq.x)
      for (i in 1:ncx){
         x0<-uniq.x[i,]
         tt<-c(abs(idxall-t(matrix(x0,p,n)))%*%matrix(1,p,1))
         idx[tt==0]<-i
      }
   }
   else { # when x is a vector
      xx<-rank(x)/n
      idxall<-idx<-(1:n)*0
      for (i in 1:n){
         idxall[i]<-length(idx0[idx0<xx[i]])+1
      }
      uniq.x<-unique(idxall)
      ncx<-length(uniq.x)
      for (i in 1:ncx){
         x0<-uniq.x[i]
         idx[idxall==x0]<-i
      }
   }

   # denominator
   den<-prod(table(y)/n)

   # numerator
   ny.x<-tapply(idx,idx,length)
   pyi<-tapply(y,idx,mean)
   num<-sum(pyi*(1-pyi)*ny.x/n)

   # integrate
   out<-1-num/den
   return(out)
}


dCvMn<-function(x,y,mx){ # m=number of bins
                         # y is a vector but x can be a nxp matrix
                         # only when y is continuous variable
   n<-length(y)
```

```
z<-sort(unique(y))
nz<-length(z)

# create the labels
idx0<-c(1:(mx-1))/mx

if (length(c(x))>n){ # when X is a matrix
   p<-ncol(x)
   xx<-x*0
   for (i in 1:p){
      xx[,i]<-rank(x[,i])
   }
   xx<-xx/n

   idxall<-x*0
   idx<-(1:n)*0
   for (i in 1:n){
      for (j in 1:p){
         idxall[i,j]<-length(idx0[idx0<xx[i,j]])+1
      }
   }
   uniq.x<-unique.array(idxall)
   ncx<-nrow(uniq.x)
   for (i in 1:ncx){
      x0<-uniq.x[i,]
      tt<-c(abs(idxall-t(matrix(x0,p,n)))%*%matrix(1,p,1))
      idx[tt==0]<-i
   }
}
else { # when x is a vector
   xx<-rank(x)/n
   idxall<-idx<-(1:n)*0
   for (i in 1:n){
      idxall[i]<-length(idx0[idx0<xx[i]])+1
   }
   uniq.x<-unique(idxall)
   ncx<-length(uniq.x)
   for (i in 1:ncx){
      x0<-uniq.x[i]
      idx[idxall==x0]<-i
   }
}

# denominator
i<-1:n
den<-1/6-1/6/n/n
```

```
   # numerator
   idx<-idx[order(y)]
   cx<-tabulate(idx)
   num<-0
   for (k in 1:ncx){
      fs<-sum(cumsum((idx==k)/cx[k]-1/n)^2)
      num<-num+fs*cx[k]/n/n
   }

   out<-num/den
   return(out)
}
```

VITA

I, Li-yu Daisy Liu, was born in Taichung, Taiwan, in November 7th, 1976. I received a B.S. degree in 1998 and a M.S. degree in 2000 both in agronomy from National Taiwan University, Taiwan. In 2000, I came to Texas A&M University to pursue a Ph.D. in statistics. My recent research interests include bioinformatics and feature selection. My permanent address is as follow,

Li-yu Daisy Liu
117 Chung-Jen Street,
Taichung 403, Taiwan

or, I can be reached at liyul@ms24.url.com.tw.